
Article:
Chen, Jingzhi, Cai, Charlie, faff, robert et al. (2022) Nonlinear limits to arbitrage. The Journal of Futures Markets. pp. 1084-1113. ISSN: 1096-9934

https://doi.org/10.1002/fut.22320

Received: 21 July 2020 | Accepted: 3 February 2022

DOI: 10.1002/fut.22320

RESEARCH ARTICLE

THE JOURNAL OF FUTURES MARKETS WILEY

Nonlinear limits to arbitrage

Jingzhi Chen[1] | **Charlie X. Cai**[2] | **Robert Faff**[3] | **Yongcheol Shin**[4]

[1]Sun Yat-Sen Business School, Sun Yat-Sen University, Guangzhou, China

[2]University of Liverpool Management School, University of Liverpool, Liverpool, UK

[3]Bond Business School, Bond University, Gold Coast, Australia

[4]Department of Economics and Related Studies, University of York, York, UK

Correspondence
Charlie X. Cai, University of Liverpool Management School, University of Liverpool, Liverpool L69 7ZH, UK.
Email: x.cai7@liverpool.ac.uk

Funding information
Economic and Social Research Council, Grant/Award Number: ES/S010238/1

Abstract

We study the nonlinear limits to arbitrage in a model. When mispricing is small, arbitrage activity increases with mispricing because of the higher cost-adjusted return. However, at high levels of mispricing, arbitrageurs are deterred by larger mispricing as funding constraints become more binding. Testing the model predictions on the index spot-futures arbitrage with a Markov-switching model, we document an inverse U-shaped relationship between mispricing and arbitrage activity. The extreme regime is with the largest mispricing but least arbitrage activity, and coincides with the market turmoil, suggesting that funding constraints become the main driver behind the limit to arbitrage.

KEYWORDS
index arbitrage, limits to arbitrage, Markov-switching GECM, mispricing correction, noise momentum

1 | INTRODUCTION

Arbitrageurs aggressively search for mispricing opportunities, which ensures that mispricing is short lived. However, arbitrage is far from a free lunch in practise. Extent finance literature has long documented that arbitrage activity is impeded by the market frictions, leading to mispricing and resource misallocations (Gromb & Vayanos, 2010). Meanwhile, larger mispricing may affect the perception on arbitrage frictions inversely, and, in turn, trigger arbitrage trades. The latter idea draws little attention in the literature, but is of great importance in understanding the complex joint determination between mispricing, arbitrage friction, and arbitrage activity.

There are two distinct and countervailing views of what limits arbitrage: arbitrage costs and funding constraints. On the one hand, previous studies (e.g., Bai & Collin-Dufresne, 2019; Gyntelberg et al., 2017; Roll et al., 2007) suggest that conducting arbitrage trade is costly and risky (e.g., market illiquidity, transaction cost, and compensation for risk). In this case, arbitrageurs are willing to exploit the mispricing only when it exceeds a certain threshold that reflects the cost of conducting the arbitrage trade. By allowing for heterogeneous arbitrage costs, a wider mispricing will trigger more aggressive arbitrage activity since it provides a higher cost/risk-adjusted return. We call it *the positive capital allocation effect*.

On the other hand, various studies build on the idea that conducting arbitrage trade requires funding, and document the importance of funding constraints in limiting arbitrage activity. The slow-moving capital hypothesis

posits that severe and prolonged mispricing especially during times of market turmoil is mainly due to the tightening of funding constraints (Acharya et al., 2010; Akbas et al., 2015, 2016; Duffie, 2010; Garleanu & Pedersen, 2011; Gromb & Vayanos, 2002; Karnaukh et al., 2015; Mitchell et al., 2007; Mitchell & Pulvino, 2012; Shleifer & Vishny, 1997). Furthermore, Brunnermeier and Pedersen (2009) suggest that larger mispricing can exaggerate expectations on future volatility, which tightens the funding constraint. In this case arbitrage activity is rather deterred in the presence of larger mispricing. We call this *the negative funding constraint effect*.

The two sources of arbitrage frictions drive opposite predictions of how arbitrageurs will respond to mispricing. The former view has been examined empirically through threshold regression models (Dwyer et al., 1996; Martens et al., 1998; Tse, 2001), while the effect of funding constraints has been studied mainly through arbitrage activity (Cielinska et al., 2017), arbitrage capital flow (Akbas et al., 2015, 2016), and violations from no-arbitrage relations (Fontaine & Garcia, 2011b; Garleanu & Pedersen, 2011). Up to our knowledge, however, there is no single study in the literature to analyze the combined impact of these two frictions on how arbitrage activity responds to mispricing, theoretically or empirically. In this paper, we address this long-standing but important knowledge deficit in the limits to arbitrage literature.

At its foundation, our empirical setup follows the standard multiperiod model of Shleifer and Vishny (1997, henceforth SV). To explicitly analyze arbitrage activities in the multiperiod setting, we also follow Cai et al. (2018, henceforth CFS) and introduce two important arbitrage parameters: the first is the initial mispricing correction parameter (κ), which measures the proportion of immediate mispricing correction achieved by arbitrageurs, and the second is the subsequent noise momentum parameter (λ), which captures the persistence of the unarbitraged pricing errors into the next period.[1] Specifically, we investigate the interaction between arbitrage costs and funding constraints through the interplay between these two parameters κ and λ with respect to the size of mispricing. When the mispricing error is small and funding is relatively ample, arbitrageurs strategically limit their investment due to concerns of arbitrage risk. In this case the arbitrage activity will intensify with the size of mispricing such that the capital allocation effect prevails. On the other hand, extremely large mispricing is likely to make funding constraints more binding. Beyond some thresholds, arbitrage activity declines with the size of mispricing because of the associated funding liquidity scarcity, suggesting that the funding constraint effect becomes the dominant driver. Accordingly, combining the two countervailing effects, the model predicts that the overall arbitrage activity displays an inverse U-shape against the size of mispricing error due to the exchange of dominance between arbitrage costs and funding constraints in limiting arbitrage.

In our empirical analysis, we apply the GECM with the Markov-switching extension (MS-GECM) to the S&P 500 index spot and futures markets over the period 1986–2015.[2] The construction of GECM, advanced by CFS, captures how arbitrage activity (both mispricing correction and noise momentum) respond to past observable mispricing, which provides a great tool to test our model predictions.[3] We find strong evidence in favor of regime-dependent nonlinear limits to arbitrage. In particular, we can identify three distinct regimes: a normal market state with a small mispricing error and low mispricing volatility, a transition market state with both medium mispricing error and volatility, and an extreme market state with a large mispricing error and high mispricing volatility. We observe a relatively low mispricing correction in the normal state, but a dramatic increase during the transition state. This suggests that arbitrage activity tends to intensify with the size of mispricing error when the mispricing level increases from low to medium. In contrast, the mispricing correction is the lowest during the extreme state. This suggests that when mispricing increases from a medium to a high level, arbitrageurs are less capable to raise external funds due to the tightening funding constraints even when the arbitrage opportunity is at its best. These extreme periods coincide with the market turmoils in the years 1987, 1998, 2001, and 2008, which provides empirical supports to the existing studies (e.g., Brunnermeier & Pedersen, 2009) documenting that the amplification effect attributed to funding illiquidity significantly jeopardizes market resiliency. Overall, arbitrage activity displays an inverse U-shape against the magnitude of mispricing errors.

To verify whether our estimation results meaningfully capture variations in the tightness of funding constraints, we examine the potential linkages between the three hidden market states and various observable measures of the funding

[1]To capture such multiperiod arbitrage activities, CFS develop a generalized error correction model (GECM) and estimate both parameters. Applying the model to a wide range of international spot–futures market pairs, CFS document pervasive evidence of noise momentum around the world.

[2]As a robustness check, we provide the estimation results over a shorter sample (1990–2015) and results employing other S&P 500 futures contracts (e.g., 6 and 9 months to maturity), which are elaborated in the appendix.

[3]In the extent literature with empirical applications using the error correction model (e.g., Balke & Fomby, 1997; Dwyer et al., 1996; Gyntelberg et al., 2017; Martens et al., 1998; Tao & Green, 2013; Theissen, 2012; Tse, 2001), past mispricing is often treated as an exogenous state variable that determines the arbitrage activity.

illiquidity.[4] Our analyses show that funding constraints keep tightening monotonically from the normal to the extreme states. The data also document the flight-to-quality/safety phenomenon such that fund flows into passive index funds decrease from normal to transition states but increase from transition to extreme states. Overall, extreme states capture a period of considerable market stress. From the arbitrageur perspective, the extreme state presents a "cocktail" of good and bad phenomena. On the positive side, it entails large mispricing errors and higher valuation uncertainty—thus presenting arbitrageurs with more profitable opportunities to exploit. On the negative side, arbitrageurs tend to face more binding funding constraints, which make them unable to capitalize on the opportunities.

Our study differs from other related studies on the limits to arbitrage and contributes to the literature as follows. First, we show that considering solely the arbitrage costs as the dominated friction is likely to obtain flawed prediction on arbitrage activity, that is, arbitrage activity is monotonically increasing with the size of mispricing. We enhance the analysis by integrating both arbitrage costs and funding constraints explanations of the limits to arbitrage in a unified framework. Our paper is the first attempt to establish that the combination of these two frictions will generate different predictions of how arbitrageurs respond to the changes in mispricing. Specifically, in the presence of large mispricing, arbitrage activity is deterred by larger mispricing due to more binding funding constraints. We denote it as the nonlinear limits to arbitrage.

Second, our empirical approach embeds the GECM advanced by CFS within the Markov-switching model, and thus contributes to the empirical literature by developing a matching empirical model that can test the validity of the nonlinear limits to arbitrage. The traditional threshold ECM (Dwyer et al., 1996; Martens et al., 1998; Tse, 2001) is insufficient to capture the nonlinear limits to arbitrage, especially in the presence of large mispricing. Our MS-GECM offers two merits. First, it embeds the multiperiod arbitrage activities proposed by CFS, which decomposes the overall speed of price adjustment into two components, that is, initial mispricing correction and subsequent noise momentum, and allows for more insightful understanding on the impact of mispricing. Second, the Markov-switching approach avoids arbitrary estimations of the threshold variable, and endogenously identifies the stress periods under which funding constraints are binding, without setting the exogenously defined crisis dates. In addition, our approach can be applied to a broad range of financial data with prices and fundamental measures (e.g., the price–dividend relationship and cross-listing and commodity contracts in different markets).

Third, it offers new insight into the relationship between funding constraints and the multiperiod arbitrage activities. The effect of funding constraints has been studied in the literature mainly through the proxies of arbitrage activity (Cielinska et al., 2017), the size of arbitrage violations (Fontaine & Garcia, 2011b; Frazzini & Pedersen, 2014; Garleanu & Pedersen, 2011) and market liquidity (Nagel, 2012; Schuster & Uhrig-Homburg, 2015). Our paper attempts to study the funding constraint effect directly through examining the relationship between arbitrage activity and the size of mispricing. A similar idea was proposed by Duffie (2010), who suggests that price reversal (measured by the speed of adjustment) provides insights regarding the arbitrage frictions borne by arbitrageurs.

Last but not least, our analysis enhances the understanding of the joint determination among mispricing, arbitrage friction, and arbitrage activity. On the one hand, arbitrage frictions, such as hedge fund flows in Akbas et al. (2015, 2016), arbitrage costs in Bai and Collin-Dufresne (2019), banking regulation in Du et al. (2018), and financial relationship in Kondo and Papanikolaou (2015), tend to limit arbitrage trade and induce wider mispricing, which is consistent with the theory of limits to arbitrage. Our time-series analysis along with the MS-GECM approach, on the other hand, studies inversely how arbitrage activity is affected by mispricing, which has received less attention in the literature. We are able to capture the nonlinear nature of limits to arbitrage, such that arbitrage activity can be triggered or deterred by larger mispricing, depending on the dominance of capital allocation or funding constraint effect. Such nonlinearity is crucial in understanding the financial market stability. The stress periods we captured in the MS-GECM coincide with the market turmoils that have extremely large mispricing but least arbitrage activity, which provides support to the funding liquidity theory in Brunnermeier and Pedersen (2009).

The remainder of our paper is organized as follows. In Section 2 we present a theoretical framework by combining the models of Shleifer and Vishny (1997) and CFS. In Section 3 we develop the main predictions on the nonlinear relation between the size of mispricing and arbitrage activities. In Section 4 we develop an MS-GECM empirical framework, designed to capture various predictions derived from Section 3. Section 4.3 presents the main empirical results for the S&P 500 index spot and futures markets. In Section 5 we make concluding remarks.

[4]For funding conditions and capital structure we use hedge and mutual fund flows, growth rate of total financial assets, financial sector leverage, and broker–dealer leverage; for illiquidity we use the Amihud (2002) illiquidity measure of the spot index, Treasury security-based funding illiquidity of Fontaine and Garcia (2011b) and The Treasury–Eurodollar (TED) spread.

2 | THEORY AND PREDICTIONS

2.1 | The model

We begin with an introduction to a range of basic concepts in line with the SV setup of limits to arbitrage. There is one asset with fundamental value V, traded by three types of market participants: noise traders, arbitrageurs, and fund investors, in three periods, $t = 1, 2, 3$, at price P_t. Noise traders arrive in period t with the demand of $(V - S_t)/P_t$, where S_t represents the extent to which noise traders in aggregate undervalue the asset price relative to its fundamental value, V. In particular, S_1 is observable to arbitrageurs; S_2 is allowed to be stochastic, taking a value of 0 (the "good" state pertains) with probability $1 - q$, or $S_2 = S_{2b} > S_1$ (the "bad" state pertains) with probability q; $S_3 = 0$, such that price converges to fundamental value in period 3, $P_3 = V$.

Observing the price discrepancy, rational and risk-neutral arbitrageurs accumulate funding resources, F_t, to explore mispricing opportunity with the demand of $\beta_t F_t/P_t$, where $0 < \beta_t \leq 1$ implies the fraction of funding to invest in the asset and $1 - \beta_t$ will invest in cash at zero interest. Since the asset is assumed to have unit supply, market clearing implies that the price of the asset is $P_t = V - S_t + \beta_t F_t$. In period 1, F_1 is exogenous given as the initial arbitrage funds.[5] We follow the maintained assumption that $F_1 < S_1$, such that arbitrageurs accumulate limited funding in period 1. In period 2, F_2 is determined endogenously by past performance, such that

$$F_{2g} = F_1\left[1 + \alpha\beta_1\left(\frac{P_{2g}}{P_1} - 1\right)\right] \quad \text{and} \quad F_{2b} = F_1\left[1 + \alpha\beta_1\left(\frac{P_{2b}}{P_1} - 1\right)\right], \tag{1}$$

where

$$P_1 = V - S_1 + \beta_1 F_1, \tag{2}$$

$$P_{2g} = V, \quad P_{2b} = V - S_{2b} + F_{2b}. \tag{3}$$

F_{2g} and P_{2g} (F_{2b} and P_{2b}) are the funding resources and the asset price in period 2 under good (bad) state; in the bad state, $\beta_2 = 1$ as price will converge to the fundamental value in period 3. $\alpha > 1$ captures the sensitivity of fund flows to past performance. If $P_2 < P_1$, arbitrageurs lose more funds than the negative return, since investors will withdraw funds based on the poor performance (this is described as the performance-based arbitrage in SV). To keep the analysis tractable and reasonable, we make a technical assumption on α to avoid overly extreme sensitivity, and derive the stability condition as follows:

$$\alpha < \alpha^* = \frac{V - S_1 + F_1}{S_{2b} - S_1 + F_1}. \tag{4}$$

We elaborate the assumption on α in the appendix.

2.2 | The equilibrium

Under this model setup, arbitrageurs actively choose their optimal investment strategy, denoted $\hat{\beta}_1$, subject to funding constraint, $0 < \beta_1 \leq 1$, so as to maximize their wealth in period 3,

$$\mathrm{E}(F_3) = (1 - q)F_{2g} + q\frac{V}{P_{2b}}F_{2b}.$$

[5]Following Shleifer and Vishny (1997) and Gromb and Vayanos (2002, 2010), we treat the initial mispricing (S_1) and the initial wealth (F_1) of arbitrageurs as exogenous, both of which are observable to arbitrageurs and affect their decision making. These assumptions seem realistic for us to understand the short-run arbitrage activity.

We obtain the following first-order condition (FOC):

$$\left(\frac{V}{P_1}-1\right)-q\left(\frac{V}{P_{2b}}-1\right)\geq 0. \tag{5}$$

$\frac{V}{P_1}-1$ indicates the return of investing in period 1, while $q\left(\frac{V}{P_{2b}}-1\right)$ is the expected return of waiting in period 1 but investing in period 2. If the return of investment in periods 1 and 2 is indifferent, the equality holds and the partial investment equilibrium ($\hat{\beta}_1<1$) is achieved. On the other hand, if investment in period 1 provides a higher return, then the inequality holds such that the full investment strategy ($\hat{\beta}_1=1$) is optimal.

It is easily seen that given the limited funding assumption ($F_1<S_1$), we must have $P_1<V$, such that mispricing in period 1 is not fully arbitraged away. Hence arbitrageurs enjoy a positive return and fund augmentation if the good state occurs ($P_1<V=P_{2g}$). Given that $0<q<1$, we have $P_{2b}<P_1$ in the partial investment equilibrium from the FOC. In the full investment equilibrium, we also have $P_{2b}<P_1$ given the stability condition in Equation (4). This can be seen by rewriting F_{2b} as

$$F_{2b}=F_1-\alpha F_1\left(\frac{S_{2b}-S_1}{V-S_1+F_1-\alpha F_1}\right)$$

with $\hat{\beta}_1=1$. The stability condition ensures that $V-S_1+F_1-\alpha F_1>0$, which results in $F_{2b}<F_1$ and $P_{2b}<P_1$. Mispricing deepens if the bad state occurs and arbitrageurs lose funds from the investment and investors' withdrawals. All in all, for arbitrageurs with limited funding, these mispricing opportunities are not risk-free.

In the full investment equilibrium (referred to as the extreme circumstance), SV find that arbitrageurs are less aggressive when mispricing opportunities are the best. More precisely, they are forced to liquidate in the bad state, that is, holding fewer shares of the asset, that is, $F_{2b}/P_{2b}<\hat{\beta}_1F_1/P_1$. It is worth noticing that full investment is a sufficient but not a necessary condition for forced liquidation, which is formally summarized in Lemma 1.

Lemma 1. *Consider the three-period model setup described in Section 2. Arbitrageurs are forced to liquidate their holdings in the bad state, that is, $F_{2b}/P_{2b}<\hat{\beta}_1F_1/P_1$, when the optimal strategy $\hat{\beta}_1$ surpasses a threshold $\beta_1^{liq}=\frac{V-S_1}{V-S_1+(\alpha-1)(S_{2b}-S_1)}$.*

Lemma 1 suggests that forced liquidation arises as a result of the performance-based arbitrage setting. For $\alpha>1$, we have $\beta_1^{liq}<1$, and forced liquidation appears in the bad state as $\hat{\beta}_1>\beta_1^{liq}$. The threshold β_1^{liq}, however, is very close to 1,[6] which implies that arbitrageurs are forced to liquidate when they adopt or very close to adopt the full investment strategy in period 1. In contrast, when arbitrageurs adopt the partial investment strategy, forced liquidation is unlikely to take place. In this regard, arbitrage activity in period 1 also reveals vital information about the market resiliency in the future.

2.3 | Arbitrage activity

As arbitrage activities in periods 1 and 2 are both informative, we now extend the SV analysis by formally defining and analyzing the arbitrage activities in both periods. CFS introduce the concept of the initial mispricing correction and the subsequent mispricing persistence (called "noise momentum") that characterize the arbitrage impact on both immediate and subsequent price movements as well as the duration of price adjustment. CFS define the initial mispricing correction as

$$K=\frac{D_1}{S_1}=\frac{\beta_1F_1}{S_1}, \tag{6}$$

[6]We can illustrate this with the numerical example in SV. Given that $V=1$, $S_1=0.3$, $S_{2b}=0.4$, $\alpha=1.2$, $F_1=0.2$, we derive the threshold for forced liquidation: $\beta_1^{liq}=0.972$. If the optimal strategy $\hat{\beta}_1$ surpasses 0.972, forced liquidation occurs in the bad state.

which is designed to capture the proportion of mispricing correction achieved by arbitrageurs in period 1. More importantly, CFS define subsequent noise momentum as

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

which captures the degree of mispricing error persistence into the next period. In particular, $V - P_1$ ($V - P_2$) represents the pricing error, which has not been arbitraged away after period 1 (2) trading. The ratio Λ represents the degree of unarbitraged error persistence. Λ is one if $P_2 = P_1$, such that all mispricing in period 1 persists to period 2. Λ becomes zero if P_2 recovers to fundamental, $V = P_2$, such that none of the unarbitraged mispricings persists. CFS demonstrate that the inclusion of noise momentum in the GECM provides a useful framework for analyzing the asset pricing dynamics and overall price adjustment process.[7]

Next, we derive the period-1 expectations of these two parameters with respect to q as follows:

$$\kappa = E_q(K) = \hat{\beta}_1 \frac{F_1}{S_1}, \qquad \lambda = E_q(\Lambda) = q\frac{V - P_{2b}}{V - P_1}, \tag{8}$$

where $\hat{\beta}_1$ is the optimal strategy achieved by the FOC in Equation (5). Notice that κ and λ are closely connected to the empirical model specification defined in Equation (10) (see CFS for more details). It is clear from Equation (8) that both κ and λ are below unity,[8] when rational arbitrageurs choose the equilibrium investment strategy.

Notice that the initial mispricing correction, κ, is the product of $\hat{\beta}_1$ and the ratio F_1/S_1. The equilibrium investment strategy $\hat{\beta}_1$ captures the strategic response of arbitrageurs to the risky arbitrage opportunity: whether to invest in period 1 or to wait and invest in period 2. It represents the willingness of arbitrageurs to engage in arbitrage activity. The term F_1/S_1 is the ratio of available funding over mispricing, which captures the arbitrageurs' funding condition. It reflects the arbitrageurs' capability of conducting arbitrage activity.

While the mispricing correction parameter, κ measures the immediate arbitrage effect, the noise momentum parameter, λ captures the subsequent price recovery. We find from Equation (8) that λ is expressed as a product of the probability, q of the bad state occurring in period 2, and the degree of error persistence $\left(\frac{V - P_{2b}}{V - P_1}\right)$ in the bad state. λ becomes higher when P_{2b} is lower than P_1, which renders the arbitrage losses higher under the bad state. Simultaneously, subsequent price adjustments become more volatile, which deters arbitrage activity.

It is easily seen that the overall speed of price adjustment is determined by both parameters. In Section 4 we will show that the speed of adjustment is positively associated with κ, but negatively with λ. This implies that the higher mispricing correction improves the price adjustment, while the higher noise momentum tends to slow it down.

3 | MISPRICING AND ARBITRAGE ACTIVITY

We now investigate how arbitrage activities respond with respect to changes in mispricing under the partial and full investment equilibria. Such an interplay between the two key arbitrage parameters, κ and λ, would reveal insights on the limits of arbitrage and the market resiliency.[9] We first derive the theoretical results while holding

[7]CFS provide a strong empirical support that the traditional ECM is misspecified in the presence of nonzero noise momentum by investigating 26 index futures relationships around the globe.

[8]Rewriting the FOC in Equation (5) as

$$\frac{V - P_1}{P_1} \geq q\left(\frac{V - P_{2b}}{P_{2b}}\right),$$

it is easily seen that

$$\lambda = q\frac{V - P_{2b}}{V - P_1} \leq \frac{P_{2b}}{P_1} < 1.$$

[9]From a long-run perspective, mispricing and arbitrage activity are determined jointly (Stein, 2009). Our paper, however, focuses on the short-run arbitrage activity in response to mispricing. The initial mispricing (S_1) is treated as exogenous: a variable that is observable to arbitrageurs and affects their decision making. These decisions, whether full investment is adopted or not, will determine the persistent mispricing and market resiliency in the future.

the probability q constant, and next we relax this assumption by allowing q to be uniformly distributed over the unit interval.

3.1 | The main theoretical predictions

We first summarize the impact of S_1, that is, the size of mispricing before arbitrageurs enter the market, on the period-1 strategy, $\hat{\beta}_1$ and the pricing efficiency in periods 1 and 2 in Lemma 2.

Lemma 2. *Consider the three-period model setup described in Section 2. Under the partial investment equilibrium, we have* $\frac{\partial \hat{\beta}_1}{\partial S_1} > 0$, $-1 < \frac{\partial P_1}{\partial S_1} < 0$, *and* $\frac{\partial P_{2b}}{\partial S_1} < 0$. *Under the full investment equilibrium,* $\frac{\partial P_1}{\partial S_1} = -1$ *and* $\frac{\partial P_{2b}}{\partial S_1} > 0$.

It is clear from Lemma 2 that the positive capital allocation effect appears under the partial investment equilibrium. As mispricing S_1 rises, arbitrageurs are willing to allocate more resources to correct the mispricing (i.e., $d\hat{\beta}_1/dS_1 > 0$). At the same time we also observe that the funding liquidity condition becomes worse (i.e., a lower F_1/S_1) with S_1, which is referred to as the negative funding constraint effect. These two contradicting effects will shape the initial mispricing correction, κ, in response to mispricing.

Consider the impact of S_1 on P_{2b},[10] which is missing in the SV analysis but has important implication when deriving the predictions on λ. Due to the capital allocation effect, arbitrageurs hold less cash in period 1 with larger S_1, which reduces their ability to bear against mispricing in the bad state of period 2. So we have $\frac{\partial P_{2b}}{\partial S_1} < 0$ under the partial investment equilibrium. Under the full investment equilibrium, however, the binding funding constraint prevents further investment in period 1 when S_1 increases. It severely distorts the pricing efficiency in period 1 $\left(\frac{\partial P_1}{\partial S_1} = -1\right)$, but improves the pricing efficiency in the bad state of period 2 as potential loss is relatively reduced and more funding is preserved. Namely, we have $\frac{\partial P_{2b}}{\partial S_1} > 0$ under the full investment equilibrium.

Next, we provide the first main theoretical results in Proposition 1.

Proposition 1. *Consider the three-period model setup described in Section 2. The impacts of initial mispricing error* (S_1) *on the arbitrage activities* $(\kappa$ *and* $\lambda)$ *are given by*

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

Furthermore, we find that the impact on noise momentum (λ) *is stronger in the full investment equilibrium than in the partial investment equilibrium:*

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

Proposition 1 shows that the positive capital allocation effect dominates the negative funding constraint effect under the partial investment equilibrium (i.e., $\partial\kappa/\partial S_1 > 0$). This is consistent with the earlier studies, documenting that larger mispricing errors induce a greater mispricing correction and faster speed of adjustment.[11] In contrast, when the funding constraint binds ($\hat{\beta}_1 = 1$), the initial mispricing correction is determined mainly by the negative funding constraint effect, since the change in mispricing does no longer have any effect (i.e., $\partial\hat{\beta}_1/\partial S_1 = 0$). As S_1 grows,

[10]As discussed in SV, P_1 tends to drop with S_1 under both equilibria, implying that arbitrageurs ability to bear against mispricing is limited.
[11]See the threshold ECM (Dwyer et al., 1996; Martens et al., 1998) and the smooth transition model (Gallagher & Taylor, 2001; Tse, 2001) for such evidence.

arbitrageurs will suffer from more deteriorating funding conditions, which forces them to disengage in arbitrage activity (i.e., $\partial\kappa/\partial S_1 < 0$).

Furthermore, Proposition 1 has additional implications about the impacts on the noise momentum parameter. Consider first the partial investment equilibrium, where arbitrageurs are likely to expect the mispricing opportunity in the bad state ($V - P_{2b}$) being relatively large. This suggests that we observe a relatively high level of λ. As already described in Lemma 2, the capital allocation effect dominates, which induces $\frac{\partial P_1}{\partial S_1}, \frac{\partial P_{2b}}{\partial S_1} < 0$. Combined together, we expect that the negative impact of S_1 on λ is rather marginal. Next, turn to the full investment case. We now have $\frac{\partial P_{2b}}{\partial S_1} > 0$ due to the binding funding constraint, which renders λ declining sharply with S_1.

The results in Proposition 1 under the full investment equilibrium provide support to the predictions in Stein (2009), who suggests that the impact of constraints on arbitrage capital is a double-edged sword. On the one hand, the funding constraint reduces the initial arbitrage activity significantly (i.e., lower κ). On the other hand, it improves the subsequent pricing efficiency and avoids the potential crashes in the future (i.e., lower λ).

From Proposition 1 we can also derive the implications on the overall speed of price adjustment. In the partial investment equilibrium, κ rises and λ marginally falls with S_1. In this state, the overall speed of adjustment (SOA) improves with S_1. But, when the funding constraint becomes binding, we find that κ rather drops with S_1. At the same time, λ falls sharply with S_1. Due to these opposite impacts, the SOA is uncertain with S_1 and should be determined empirically.

So far we investigate how arbitrage activities respond to mispricing while holding the probability q constant. However, q is far from constant over time in reality. Hence we now extend the SV setup by allowing q to be time varying. Specifically, we assume, for simplicity, that q is uniformly distributed over $(0, 1)$.[12] SV introduce a threshold probability, q^*, for the given model parameters, V, S_1, S_{2b}, F_1, and α. If $q < q^*$ (i.e., the probability of the mispricing deepening in period 2 is relatively low), then arbitrageurs will adopt the full investment strategy in period 1. Alternatively, if $q > q^*$, arbitrageurs will defer some of their investment. We now provide the formal link between the size of mispricing and the threshold parameter, q^* that crucially determines the arbitrageur's investment strategy.

Proposition 2. *Consider the three-period model setup described in Section 2. As the mispricing error* (S_1) *rises, the threshold probability,* q^* *also rises* (*i.e.*, $\frac{\partial q^*}{\partial S_1} > 0$). *In the two extreme cases, we have* $\lim_{S_1 \to F_1} q^* = 0$ *and* $\lim_{S_1 \to S_{2b}} q^* = 1$.

Proposition 2 is intuitive. $\lim_{S_1 \to F_1} q^* = 0$ implies that with small initial mispricing, the full investment strategy is not optimal for any q, since we always have $q > q^* = 0$. As mispricing grows larger, funding condition becomes worse. Arbitrageurs are more likely to be fully invested, where the funding constraint binds, since q^* is monotonically rising with S_1. For extremely large mispricing, the full investment equilibrium is inevitable, as $q < q^* = 1$.

Given that $q \sim U(0, 1)$, we redefine the period-1 expectation of the arbitrage activities as

$$\kappa_q = \sum_{j \in J} P(q = j)\kappa(q = j), \quad \lambda_q = \sum_{j \in J} P(q = j)\lambda(q = j), \tag{9}$$

where J denotes the set of all real numbers within the unit interval, $P(q = j)$ is the probability of $q = j$ and $\kappa(q = j)$ and $\lambda(q = j)$ are the initial mispricing correction and noise momentum for $q = j$. Now, we provide the second main theoretical results in Proposition 3.

Proposition 3. *Consider the three-period model setup described in Section 2, and assume that* q *is uniformly distributed over the unit interval and is independent of* S_1. *As mispricing* (S_1) *increases from* F_1 *to* S_{2b}, κ_q *displays an inverse U-shape pattern and* λ_q *drops monotonically.*

[12]Probability q is a given variable in the SV setup, which implies that q is independent of S_1. We maintain the assumption that the distribution of q is independent of S_1.

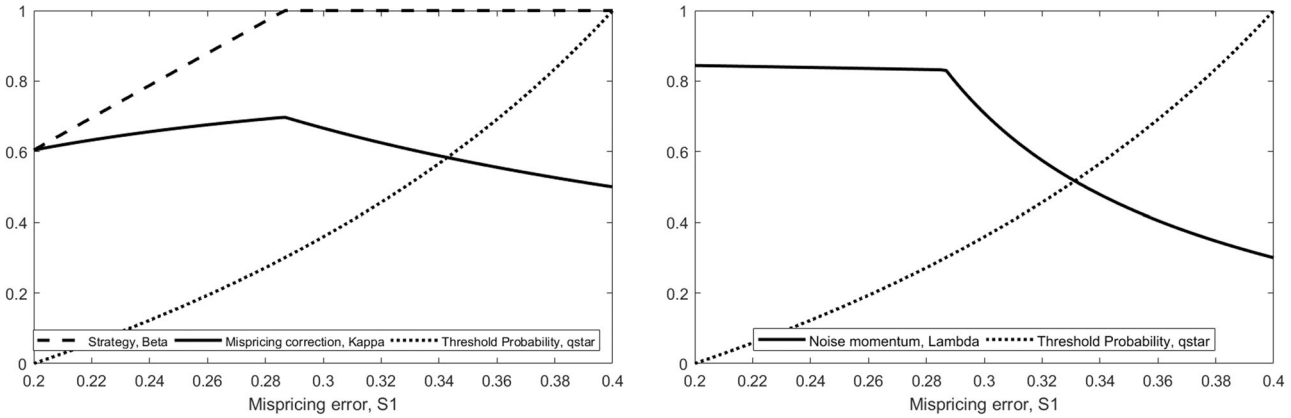

FIGURE 1 Strategic response effect on mispricing correction and noise momentum, constant q. The nonlinear impact of mispricing errors on mispricing correction (left) and noise momentum (right). It follows the numerical examples from the SV paper, such that $V = 1$, $S_{2b} = 0.4$, $q = 0.3$, $\alpha = 1.2$, $F_1 = 0.2$, and initial mispricing S_1 increases from 0.2 to 0.4. The left graph also plots the optimal investment strategy $\hat{\beta}_1$ and the threshold probability q^* of deepening mispricing, while the right graph provides an additional plot of the threshold q^*. Both graphs highlight the nonlinearity in the arbitrage parameters with respect to changes in mispricing

Proposition 3 predicts that the inverse U-shaped relationship between mispricing errors and arbitrage activities still holds, given that q is uniformly distributed. The inverse U-shape arises because the likability of a full investment equilibrium increases with S_1 (Proposition 2). The partial investment is more likely to be optimal for small S_1. Thus we observe the positive capital allocation effect on average. Once most of the possible scenarios are registered as the full investment equilibrium with higher S_1, the negative funding constraint effect inevitably appears.[13]

3.2 | Numerical analysis

We now provide a numerical analysis, confirming the nonlinear limits to arbitrage. Figure 1 illustrates how arbitrageurs' initial mispricing correction (κ) and subsequent noise momentum (λ) interplay with respect to the mispricing error under the partial and full investment equilibria, following and extending the numerical example by SV. Let $V = 1$, $S_{2b} = 0.4$, $q = 0.3$, $\alpha = 1.2$, and $F_1 = 0.2$. In the left panel, as S_1 increases from 0.2 to 0.4, the initial mispricing correction (solid line) displays an inverse U-shaped relation with respect to mispricing errors, depending on whether arbitrageurs adopt the partial or the full investment strategy. In the partial investment equilibrium where the threshold q^* lies below the probability q, the capital allocation dominates the funding constraint effect so as to improve mispricing corrections. On the other hand, only the funding constraint effect remains under the full investment equilibrium, which deters mispricing corrections. The right panel displays that the noise momentum (solid line) decreases with S_1, gradually under the partial investment strategy but rather sharply under the full investment strategy. We also notice that noise momentum is relatively high and declines mildly under the partial investment equilibrium.

Next we allow probability q to be uniformly distributed over $(0, 1)$. Figure 2 shows how arbitrageurs' initial mispricing correction (κ_q) and subsequent noise momentum (λ_q) interplay with respect to the mispricing error. Similarly, we let $V = 1$, $S_{2b} = 0.4$, $\alpha = 1.2$, $F_1 = 0.2$, and $q \sim U(0, 1)$, and we allow S_1 increases from 0.2 to 0.4. The mispricing correction κ_q (the solid line) displays an inverse U-shaped relation with respect to S_1, while the noise momentum λ_q (the dashed line) declines.

[13]We note that although we extend SV's analysis to make q a distribution but q is still independent of other model parameters. What if q is a regime-dependent variable determined by the size of mispricing S_1? There is a possibility that q increases with S_1 and we have $q > q^*$ for any S_1. The model now predicts that κ drops with S_1, such that arbitrage activity becomes less aggressive when the mispricing error is larger, that is, the negative capital allocation effect. However the result is counterintuitive. Building a new framework to model the relation between q and S_1 may provide more insights into our results. We leave it to future research.

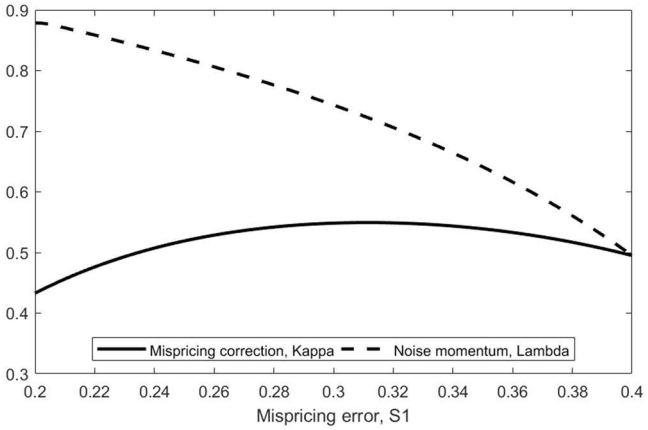

FIGURE 2 Strategic response effect on mispricing correction and noise momentum, uniformly distributed q. The nonlinear impact of mispricing errors on mispricing correction (κ_q, the solid line) and noise momentum (λ_q, the dashed line), given that q is uniformly distributed over $(0, 1)$. It follows the numerical examples from the SV paper, such that $V = 1$, $S_{2b} = 0.4$, $\alpha = 1.2$, $F_1 = 0.2$, $q \sim U(0, 1)$, and initial mispricing S_1 increases from 0.2 to 0.4. The graph again highlights the nonlinearity in the arbitrage parameters with respect to changes in mispricing

Overall, our analysis demonstrates that the mispricing correction is not always positively associated with the magnitude of mispricing error. Also, noise momentum is regime dependent. Nonlinear limits to arbitrage arises due to the interplay between arbitrage risk and funding constraints. This highlights the importance of taking into account the combined effects of arbitrage risk and funding constraints in deriving a distinctive association between mispricing and arbitrage activities under the partial and full investment equilibria.

3.3 | Hypotheses development

From the theoretical results in Proposition 3, we consider three regimes with different magnitudes of mispricing error (small, medium, and large), and denote κ^r and λ^r with $r \in \{s, m, l\}$. We then summarize three main hypotheses as follows:

Prediction 1 (Inverse U-shaped initial arbitrage activity). (*i*) $\kappa^m > \kappa^s$*; the partial investment strategy is more likely to be optimal when mispricing level is low or medium; the initial mispricing correction rises with the size of mispricing error*. (*ii*) $\kappa^l < \kappa^m$*; the full investment equilibrium tends to dominate as mispricing becomes extremely large; a further rise in mispricing error will induce a slower mispricing correction*.

Prediction 2 (Regime-dependent noise momentum). (*i*) λ^s *is relatively high, suggesting that mispricing tends to persist in period* 2 *as the size of initial mispricing error is relatively small*. (*ii*) $\lambda^m \approx \lambda^s$*; the difference between* λ^s *and* λ^m *is negligible when the mispricing level increases from low to medium*. (*iii*) λ^l *is significantly smaller than* λ^m *as mispricing increases from medium to large*.

Prediction 3 (SOA). *SOA tends to be faster with the larger mispricing error when the mispricing level increases from low to medium; In contrast, the impact of mispricing error on SOA will be uncertain when mispricing increases from medium to large*.

4 | EMPIRICAL APPLICATIONS

In this section we examine the empirical validity of the main hypotheses regarding the limits of arbitrage and its nonlinear impacts on the asset pricing dynamics developed in Section 3 by analyzing the daily S&P 500 index spot and futures contracts between June 1986 and December 2015.

4.1 | The state-dependent MS-GECM

Consider the two-period GECM advanced by CFS that captures the multiperiod (complex) arbitrage activities as follows:

$$\Delta f_t = \kappa z_{t-1} + \lambda(1 + \kappa) z_{t-2} + \delta \Delta f_t^* + \gamma \Delta f_{t-1} + u_t, \quad u_t \sim \text{iid}\left(0, \sigma_\mu^2\right), \tag{10}$$

where f_t is the (observed) spot price, f_t^* is the fundamental value for the asset, $z_t = (f_t - f_t^*)$ is the pricing error that is the short-term deviation of price from its fundamental value, Δ is the first-difference operator, and u_t is the zero-mean idiosyncratic error term with zero mean and finite variance σ_u^2, whilst $\kappa, \lambda, \delta, \gamma$ are the parameters of interest. In particular, κ captures the pricing impact of initial arbitrage activity in correcting the observable mispricing error z_{t-1} (treated as the initial mispricing S_1 in the theory), which is regarded as the percentage of mispricing correction. With $(1 + \kappa)z_{t-2}$ representing the unarbitraged mispricing error carried over from the previous period, λ measures the pricing impact of these unarbitraged errors, that is, the strength of noise momentum. These two components provide a natural framework for testing the main hypotheses regarding nonlinear limits to arbitrage. As discussed in Section 3, arbitrageurs ability to correct the initial mispricing is limited due to arbitrage risk and funding constraints, suggesting that $|\kappa|$ is below unity. The unarbitraged pricing error component persists into the next trading period, which leads to a positive λ. The overall speed of convergence to equilibrium is determined jointly by κ and λ, namely, $\kappa + \lambda(1 + \kappa)$, implying that the standard one-period ECM is likely to be biased and misleading in the case where $\lambda \neq 0$.

Moreover, δ measures the contemporaneous price reaction to the fundamental changes. Hence the recovered parameter $\omega = \delta - 1$ captures the degree of the over- or underreaction with respect to the contemporaneous fundamental changes. w tends to be different from zero unless the market is perfectly efficient, such that a one unit change in fundamentals causes one unit change in the market price instantaneously. A positive w implies that futures price overreacts to the impact of the fundamental changes irrespective of the signs. Finally γ presents the impact of the short-term momentum effect attributed to the previous price changes, the sign of which is generally ambiguous and empirically determined. The recovered parameter $\pi = -\gamma/\omega$ captures the possible feedback trading pattern, such that past market price might induce changes in fundamentals. It can be seen after rewriting Equation (10) as

$$\Delta f_t = \kappa z_{t-1} + \lambda(1 + \kappa)z_{t-2} + \Delta f_t^* + \omega e_t + u_t,$$
$$\Delta f_t^* = \pi \Delta f_{t-1} + e_t.$$

A positive (negative) π indicates positive (negative) feedback trading, such that an increase in past market price leads to fundamental growth (decline).[14]

Our key theoretical hypotheses suggest that arbitrage activity is fundamentally nonlinear, crucially depending on the magnitudes of mispricing error, as described in Section 3. By construction the linear model cannot test our hypotheses because it imposes (potentially invalid) symmetry restrictions and is thus likely to yield misleading results. Accordingly, in our empirical application, we choose to embed the GECM within the Markov-switching model popularized by Hamilton (1989), which enables us to identify the regime of binding constraints through the interplay between arbitrage parameters κ and λ. In particular, we consider a three-regime setup, which is compatible with our theoretical model that has two alternative paths to equilibrium:

$$\Delta f_t = \alpha_{R_j} + \kappa_{R_j}\hat{z}_{t-1} + \lambda_{R_j}^*\hat{z}_{t-2} + \delta_{R_j}\Delta f_t^* + \gamma_{R_j}\Delta f_{t-1} + u_{tR_j}, \quad u_{tR_j} \sim \text{iid}(0, \Sigma_{R_j}), \tag{11}$$

where f_t (f_t^*) is the natural log of the spot (fundamental) asset price, and $\{\alpha_{R_j}, \kappa_{R_j}, \lambda_{R_j}^*, \delta_{R_j}, \gamma_{R_j}\}$ are regime-dependent parameters, with $\sum_{R_j}$ being the regime-dependent covariance of the residuals. The pricing error, $\hat{z}_t$, is estimated from the following long-run equation:[15]

$$f_t = \mu + \theta f_t^* + \hat{z}_t. \tag{12}$$

The regime-specific noise momentum coefficient, λ_{R_j} can be obtained from $\lambda_{R_j}^* = \lambda_{R_j}(1 + \kappa_{R_j})$. We estimate the MS-GECM with three regimes by maximizing the log-likelihood function, where R_j is a scalar geometric ergodic Markov chain with a three-dimensional-regime space, having the following transition matrix:

[14]See CFS for detailed steps in developing the GECM that connects the theoretical model.
[15]According to the cost-of-carry model, the theoretical value of θ is 1, which is strongly supported by our empirical analysis.

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

where $P_{ij} = Pr(R_j|R_i)$ is the transition probability from States i to j. This model is designed to provide further insights into the asymmetric price discovery process in a flexible manner.

4.2 | Application to index futures

We investigate the no-arbitrage relationship between index futures contracts and the underlying spot contracts, where the fundamental value for futures contracts is implied by underlying spot price and the cost-of-carry model (Roll et al., 2007). Assuming that the risk-free rate and dividend yield are given, the fundamental value of futures contract is obtained as

$$f^*_{t,T} = i_t + (r_t - q_t)\tau_t,$$

where $f^*_{t,T}$ is the natural log of the fundamental price of the futures contract with a maturity date T at day t; i_t is the natural log of the index spot price at day t; r_t and q_t are the risk-free interest rate and dividend yield of the asset, with $\tau_t = T - t$ is the time to maturity.

We apply our approach to the daily S&P 500 index spot and futures contracts between June 1986 and December 2015.[16] The futures contracts are the most actively traded 3-month-to-maturity contracts that roll over every quarter (March, June, September, and December) into successive 3-month-to-maturity contracts. All data are sourced from Datastream (1986–2015). Our proxy for the risk-free interest rate is the US 3-month Treasury bill (T-bill) rate. Dividend yields on the indices are also collected.[17] A continuous series of the nearest-term futures contracts is constructed. These 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 for every contract to ensure correct matching of the date to maturity with the continued futures price series.[18] 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 higher in the futures contracts. On average, the basis (the log difference between futures and spot prices) is 24 basis points. After applying the cost-of-carry model, the difference between the futures price and the fair estimate ($f - f^*$) is 5.5 basis points.

4.3 | Main empirical results

The MS-GECM estimation results are reported in Table 2 with three (smoothed) regime probabilities plotted in Figure 3. We first discuss the stylized feature of three distinct regimes from Figure 3, which we call States 1–3, respectively. State 1 is the dominant market state, with the smallest mispricing error (measured as the absolute value of

[16]We did not use the early data from the period 1982 to 1986, since the estimated mispricing errors are more than double on average during this period, compared to the period over 1986–2015. The index futures contracts were first introduced in 1982, where the market had higher transaction costs and lacked index arbitrageur. Thus larger mispricing errors occurred. Errors became more stable after 1986, and fluctuated with major market events. See also Appendix B for a plot of the moving-average mispricing error through time.

[17]As a robustness check, we have also considered futures contracts with different maturities (6 and 9 months to maturity), and consider the LIBOR rate as an alternative measure for risk-free rate (van Binsbergen et al., 2019). Overall, the estimation results for the arbitrage activity are qualitatively similar to what follows, and find an inverse U-shaped limit-to-arbitrage pattern across the three market regimes. These results are available in the appendix with brief discussion.

[18]Inevitably, some error might appear when estimating the fundamental value of the futures contract. First, the measurements of risk-free rate and dividend yield can be arbitrary. Second the fact that the index spot and futures contracts are open to trade at the same time (9:30 a.m.), but close at different times (4:00 p.m. for spot and 4:15 for futures), will also introduce some measurement errors. These errors seem to be small and related to arbitrage costs only.

TABLE 1 Basic descriptive statistics

	Mean	Median	Minimum	Maximum	Std Dev
Δi	0.028	0.058	−22.833	10.957	1.169
Δf	0.028	0.062	−33.700	17.749	1.262
$f - i$	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

Note: This table reports the descriptive statistics for all variables. The sample covers the daily series of the S&P 500 index and its futures contracts over the period June 4, 1986–December 3, 2015. There are a total of 7442 observations. Δi (Δf) is the first difference of log index spot (futures) price. $f^*_{t,T} = i_t + (r_t - q_t)\tau_t$ is the fair price for the future contract, where r_t is the annualized risk-free (3-month T-bill) interest rate on investment, and q_t is the annualized dividend yield on the index. All numbers are recorded in percentage.

TABLE 2 Estimation of the Markov-switching generalized error correction model, 1986–2015

Panel A. Estimation results

	State 1		State 2		State 3		States 2–1		States 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***	332.0	1.024***	259.0	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
$\lvert z_{t-1}\rvert$	0.103		0.207		0.774					
Log-likelihood		2360.95								

Panel B. Recovered coefficients

	State 1		State 2		State 3		States 2–1		States 3–2	
	Estimate	t stat	Estimate	t stat	Estimate	t stat	Estimate	t stat	Estimate	t stat
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}
State 1_t	0.980	0.024	0.000
State 2_t	0.019	0.968	0.143
State 3_t	0.000	0.008	0.857
Ergodic	0.531	0.443	0.024

Note: This table reports the estimation of the Markov-switching GECM. The sample covers the daily series of the S&P 500 index and its futures contracts over the period June 4, 1986–December 3, 2015. There are a total of 7442 observations, of which 4070, 3222, and 148 fall into States 1, 2, and 3, respectively. Specifically, Panel A reports the estimation results for $\Delta f_t = \alpha_{R_j} + \kappa_{R_j}\hat{z}_{t-1} + \lambda^*_{R_j}\hat{z}_{t-2} + \delta_{R_j}\Delta f^*_t + \gamma_{R_j}\Delta f_{t-1} + \mu_{tR_j}$, where $\hat{z}_t$ is estimated from Equation (11), $\{\alpha_{R_j}, \delta_{R_j}, \gamma_{R_j}, \kappa_{R_j}, \lambda^*_{R_j}\}$ are regime-dependent coefficients with the covariance of the residuals (Σ_{R_j}), taking different values across the three states. Panel B reports the recovered coefficients. Specifically, $\omega_{R_j} = \delta_{R_j} - 1$, $\pi_{R_j} = -\gamma_{R_j}/\omega_{R_j}$, 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 statistics between states. For nonlinear 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. All t statistics are computed based on a numerical Hessian matrix, and ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

Abbreviations: GECM, generalized error correction model; SOA, overall speed of adjustment.

the deviations: $|z_{t-1}| = 0.103$) and mispricing volatility ($\Sigma = 0.113$) in Panel A and the highest ergodic probability (53%) in Panel C. It mainly covers three major bull markets during 1992–1995, 2003–2007, and 2012–2015, and we call it the normal market state. State 2 covers 44% of the sample, with mispricing error (0.207) and volatility (0.239) at twice the levels seen in State 1. This state corresponds to the periods 1986–1991, 1996–2002, 2009, and 2011, which are mostly the transition periods before and after the major crises. We refer to this period as the transition market state. Finally, State 3 is characterized by extremely large mispricing error (0.774) and volatility (1.077).[19] It covers only 2.4% of the sample, and coincides mostly with the stressed episodes that are captured in our sample period including the stock market crash in 1987, the Russian financial crisis in 1998, the market meltdown in 2001, and the global financial crisis in 2008. We call this the extreme market state.

Moreover, State 1 is most persistent with 98% transition probability, followed by State 2 with 97% and State 3 with 86% (see Panel C in Table 2), which suggests that State 1 (3) is the most (least) "sticky." The transition probabilities between States 1 and 3 in either direction are nil, confirming that State 2 is indeed the transition market state. The three distinct market states identified by the MS-GECM are mostly consistent with the different historic episodes we have observed during the whole sample period.

Linking the findings in Table 2 to our key hypotheses, we find that the estimated mispricing correction parameters, κ_{R_j}, are all negative and significantly less than unity in the absolute sense, in all three regimes. State 2 displays the fastest initial correction (82%), followed by State 1 (70%) and State 3 (60%). Comparing the difference between κ coefficients across the three different market conditions, it appears that arbitrageurs play a bigger role in bringing the price back to its fundamental value when switching from States 1 to 2, with the difference (–12%) being statistically significant. In contrast, the coefficient differential between States 2 and 3 becomes significantly positive (22.4%), suggesting that arbitrage activity is rather limited even though the mispricing error in State 3 is 3.7 times higher than in State 2. These findings provide strong support for Prediction 1 that initial arbitrage activity follows the inverse U-shaped pattern with respect to the size of mispricing errors.

Next, we turn to the noise momentum coefficients, λ_{R_j}, reported in Panel B. Notably, we find that the strength of noise momentum is significant and relatively high during normal and transition market states, implying that the unarbitraged error coming from the previous period is highly persistent, respectively at 82% and 92%. In contrast, noise momentum is negligible (28%) during the extreme market state, given that the coefficient is not statistically significant. The difference in λ coefficients between States 2 and 1 is insignificant and negligible, while the difference between States 3 and 2 is significant and negative (–65%). The significant drop in λ in State 3 is consistent with the impact of a binding funding constraint that improves futures pricing efficiency at the cost of distorting the current one. Overall, this finding provides support for Prediction 2, on the regime-dependent differences in noise momentum. Especially, it highlights that λ is another important parameter in characterizing the SOA process.

Combining both mispricing correction and noise momentum coefficients, we find that overall speeds of adjustment (given by $\kappa + \lambda(1 + \kappa)$) are 45%, 65%, and 60%, respectively, for the normal, transition, and extreme regimes. The SOA becomes significantly faster when the market switches from States 1 to 2 attributed to the increment of mispricing correction. The combination of initial correction and size of mispricing suggests that the capital allocation effect is the main driver behind the different adjustment speeds between these two states. In contrast, when we move from States 2 to 3, we observe that κ plays an important role in slowing down the SOA. This evidence prompts new insights into the cause(s) of a prolonged error correction process, which in previous literature is often explained by the presence of larger transaction costs (Bai & Collin-Dufresne, 2019; Gyntelberg et al., 2017; Roll et al., 2007). Consistent with the perception of binding funding constraints, we find initial correction and noise momentum are significantly lower in extreme market conditions. In other words, arbitrageurs expect initial mispricing correction to be limited but future mispricing to be relatively improved. The interplay between κ and λ against the extremely large mispricing suggests that the funding constraint effect is the main driver behind the outcome that overall arbitrage activity is significantly deterred in State 3. As such, our empirical findings provide extra empirical support for the slow-moving capital hypothesis in the literature (Brunnermeier & Pedersen, 2009; Duffie, 2010; Mitchell et al., 2007; Mitchell & Pulvino, 2012).

We also observe a range of notable findings for the other parameters in our model. First, the intercepts, α_{R_j}, in States 1 and 2 are statistically significant. A large positive intercept is found in State 2, while a negligible intercept is found in State 1—the positive sign indicating a regime in which the futures price is more bullish than the spot

[19]Notice that the average mispricing error over the whole sample period is 0.161.

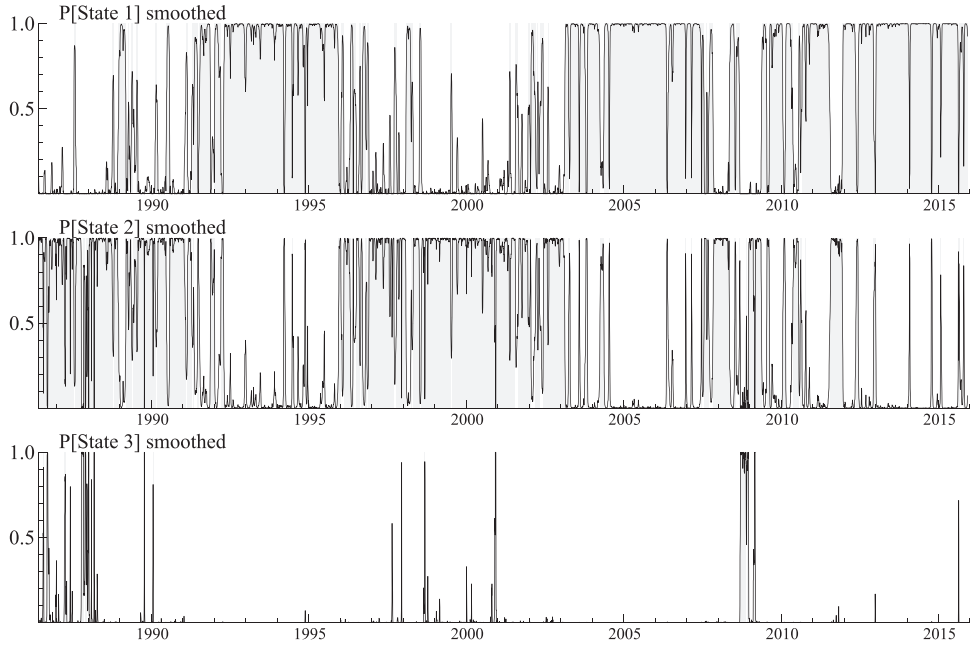

FIGURE 3 The smoothed regime probability, 1986–2015. The smoothed regime probability of being in State 1, 2, or 3. It is resulted from the estimation of MS-GECM with 3-month-to-maturity futures contracts over 1986–2015. State 1 consists of the years 1992–1995, 2003–2007, and 2012–2015, State 2 consists of the years 1986–1991, 1996–2002, 2009, and 2011, while State 3 is found in the years of 1987, 1998, 2001, and 2008. MS-GECM, Markov-switching generalized error correction model

price, other things equal. As such, this suggests that the transition state coincides with periods in which the futures market is more bullish than the spot market. On average, during this regime, there is a 1.6 basis-point daily return in the futures market, regardless of the spot market movement. However, such returns are accompanied by larger risks, as reflected by the variance of the futures return. The pricing difference between the index and futures markets under different market conditions is not directly considered in previous studies, and thus adds a new result to the literature. Second, Table 2 shows that the contemporaneous market reaction coefficient, $w_{R_j} = \delta_{R_j} - 1$, is statistically different from zero. It is small and negative in the normal state, while it is highly significant and positive in the transition and extreme states. In the case of the latter, it suggests that for a 1 percentage-point change (up or down) in the fundamental value there will be a 0.024 or 0.142 percentage-point price overreaction, respectively, in the futures market in the same direction (i.e., a 2.4 or 14 basis-point overreaction, respectively). Third, Table 2 also shows that while in the normal market state there is no feedback trading (i.e., we cannot reject $\pi = 0$), there is large, negative, and significant feedback trading in transition and extreme market conditions. Such negative feedback trading leads to more intense price reversal, which is consistent with the high volatility observed in these market conditions.

4.4 | Linking hidden states to observables

The advantage of using a Markov-switching model is the capacity to estimate the likelihood of the market being in a given latent state, which can then be examined for potential linkages to various observable economic factors. It offers an opportunity to better understand and characterize what the regimes are really capturing. In particular, we are interested in how well the model classifies the market states according to funding conditions.

Table 3 reports the mean and median statistics of monthly funding and liquidity measures in each of the three states. Months are allocated to states according to the dominant state for that month (i.e., the state with the largest number of days, or highest probability). We have 189, 160, and 6 months of observations for States 1, 2, and 3, respectively. Variable definitions are given in Appendix C. We include the Chicago Board Options Exchange (CBOE) Volatility Index (VIX) and variables of aggregated hedge and mutual fund flows, capital structure and stock, market liquidity.

TABLE 3 Linking hidden states to observables

	State 1		State 2		State 3	
Variables	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**
VIX	15.95	15.22	24.24	23.30	52.35	55.31
Fund flows						
Hedge fund flows	10.22	7.92	5.43	4.99	1.83	0.14
Active fund flows	130.72	135.69	74.04	78.42	40.40	51.07
Index fund flows	9.49	8.66	6.81	6.29	9.13	9.10
Capital structure						
Growth rate of total financial assets	0.02	0.03	0.05	0.04	−0.01	−0.03
Financial sector leverage	7.89	3.60	12.09	5.21	30.94	38.87
Broker–dealer leverage factor	43.59	42.21	39.54	39.68	61.81	59.38
Liquidity						
Amihud illiquidity of SPX 500	28.29	2.76	85.59	14.55	71.26	3.78
Treasury security-based funding illiquidity	−0.21	−0.23	0.16	0.25	1.81	1.96
TED spread	0.00	0.00	0.01	0.01	0.03	0.03

Note: This table reports the mean and median statistics of monthly funding and liquidity measures in the three regimes. Variable definitions are given in Appendix C. Months are allocated to states according to the dominant state for that month (i.e., the state with the largest number of days, or highest probability). We have 189, 160, and 6 months of observations for States 1, 2, and 3, respectively.

Abbreviations: TED, The Treasury–Eurodollar; VIX, Volatility Index.

The VIX values confirm that volatility increases from States 1 to 3, with State 3 having a value of more than double that of State 2. Fund-flow statistics suggest that funds available for arbitrageurs (hedge and active mutual funds) decrease as volatility increases from States 1 to 3. The decreasing funding flow confirms that funding constraints are more likely to be binding in the extreme state when inflows to hedge fund and active funds are at their lowest. In contrast, passive index funds received large inflows during this same extreme period, which suggests that funding constraints arise partly due to the relocation of investment to equity-focused funds, which confirms the flight-to-quality/safety phenomenon. The results for the growth rate of total financial assets and leverage factors point to a similar story, that funding constraints are most binding in State 3. Financial asset growth is at its lowest (in fact negative) while financial sector leverage and broker–dealer leverage are highest. The differences between States 1 and 2 are small relative to their differences from State 3, which confirms State 3 as the most extreme of the three regimes.

With regard to market liquidity, the Amihud illiquidity measure suggests that it plays little role in affecting arbitrageur decisions in between different market states. In general, we expect that spot market illiquidity would deter arbitrage, but we observe the contrary when considering States 2 and 3: although the spot market is more illiquid in State 2, we observe larger error correction. Bond illiquidity and TED spread as a measure of funding liquidity risk (Fontaine & Garcia, 2011b) confirm that funding liquidity risk is at its highest in State 3. The difference between States 2 and 3 is much higher than that between States 1 and 3.

Overall, our findings confirm that our regime estimations capture the variations in funding constraints and therefore support our hypothesis that funding constraints are an important driver of variation in arbitrage activity, especially during the extreme market state.

5 | CONCLUSIONS

We develop a unified approach to generate key predictions regarding the effects of capital allocation and funding constraints on the limits to arbitrage. Building on the seminal work by Shleifer and Vishny (1997) where arbitrageurs attempt to exploit mispricing while simultaneously facing arbitrage risk and funding constraints, we analyze the nonlinear impacts of mispricing on arbitrage activity. To capture the multiperiod arbitrage activity, we follow Cai et al.

(2018) and investigate both the initial mispricing correction and the subsequent noise momentum parameters, where the latter is designed to measure the persistence of uncorrected pricing errors.

In the normal state, where mispricing opportunity is small and arbitrage funds are ample, arbitrageurs are cautious due to the costs and risk of conducting the arbitrage trade. The capital allocation effect dominates, and larger mispricing induce more aggressive arbitrage activity because of the higher cost-adjusted returns. But, under the extreme state characterized by extremely large mispricing and high volatility, funding constraint is more likely to be binding. The negative funding constraint effect becomes the dominant driver in limiting arbitrage, and larger mispricing rather impeded arbitrage activity due to intensified funding scarcity. Our theory predicts that overall arbitrage activity does not rise linearly with mispricing error. Rather, the relationship tends to be regime dependent such that overall arbitrage activity displays an inverse U-shape against the size of mispricing error.

To test the validity of our theoretical predictions, we extend the multiperiod error correction model by CFS to the state-dependent Markov-switching model. Applying this model with three regimes to the S&P 500 index spot and futures markets over the period 1986–2015, we find strong evidence in favor of regime-dependent nonlinear limits to arbitrage. Furthermore, we can endogenously identify the stress periods of binding funding constraints as the years 1987, 1998, 2001, and 2008. In this regard, our study can provide both theoretical and empirical evidence that is consistent with the slow-moving capital hypothesis documented in the literature (Brunnermeier & Pedersen, 2009; Mitchell et al., 2007).

Finally, we note that our framework could easily be extended to address a range of further issues. First, our approach could be applied to explore the short-term dynamics associated with fundamental long-run co-integrating relationships (e.g., the price–dividend relationship) and the pricing dynamics between segmented markets for single assets (e.g., cross-listing and commodity contracts in different markets). Second, as our paper highlights the fruitful results of studying the limits to arbitrage via arbitrage activities, another important extension would be to analyze the cross-sectional effects of specific arbitrage impediments to arbitrage activity. Third, our paper aims to analyze arbitrage activities in response to mispricing observed exogenously. It would be interesting to see how arbitrage activities and mispricing will interact and amplify.

ACKNOWLEDGMENTS

We are grateful to Adam Golinski, Minh Nguyen, Tyler Shumway, Andy Snell, Huamao Wang, Tim Worrall and the seminar participants at University of York, University of Edinburgh, University of Kent, the 22nd Spring Meeting of Young Economists, and the 25th Annual Symposium of the Society for Nonlinear Dynamics and Econometrics for their helpful comments. Chen acknowledges financial support from Department of Economics and Related Studies, University of York and the ESRC Postdoctoral Grant (ES/S010238/1). Any errors or omissions are the responsibility of the authors. The usual disclaimer applies.

DATA AVAILABILITY STATEMENT

The data of S&P 500 index and index futures prices are available from Thomson Reuters Datastream. The data of various observable economic factors are available from Chicago Board Options Exchange, COMPUSTAT, CRSP database, Federal Reserve Economic Data library, Morningstar Direct database, the website of Tyler Muir and Jean-Sebastien Fontaine. Restrictions apply to the availability of the data sets, which were used under license for this study.

ORCID

Jingzhi Chen http://orcid.org/0000-0001-9695-0059
Charlie X. Cai http://orcid.org/0000-0003-1398-3715

REFERENCES

Acharya, V. V., Shin, H. S., & Yorulmazer, T. (2010). Crisis resolution and bank liquidity. *The Review of Financial Studies*, *24*(6), 2166–2205.

Adrian, T., Etula, E., & Muir, T. (2014). Financial intermediaries and the cross-section of asset returns. *The Journal of Finance*, *69*(6), 2557–2596.

Adrian, T., & Shin, H. S. (2010). Liquidity and leverage. *The Journal of Financial Intermediation*, *19*(3), 418–437.

Akbas, F., Armstrong, W. J., Sorescu, S., & Subrahmanyam, A. (2015). Smart money, dumb money, and capital market anomalies. *Journal of Financial Economics*, *118*(2), 355–382.

Akbas, F., Armstrong, W. J., Sorescu, S., & Subrahmanyam, A. (2016). Capital market efficiency and arbitrage efficacy. *Journal of Financial and Quantitative Analysis*, *51*(02), 387–413.

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, *5*(1), 31–56.

Ang, A., Gorovyy, S., & Van Inwegen, G. B. (2011). Hedge fund leverage. *Journal of Financial Economics*, *102*(1), 102–126.

Bai, J., & Collin-Dufresne, P. (2019). The CDS-bond basis. *Financial Management*, *48*(2), 417–439.

Balke, N. S., & Fomby, T. B. (1997). Threshold cointegration. *The International Economic Review*, *38*, 627–645.

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The Review of Financial Studies*, *22*(6), 2201–2238.

Cai, C. X., Faff, R., & Shin, Y. (2018). Noise momentum around the world. *Abacus*, *54*(1), 79–104.

Cielinska, O., Joseph, A., Shreyas, U. P., Tanner, J., & Vasios, M. (2017). *Gauging market dynamics using trade repository data: The case of the Swiss franc de-pegging* (Financial Stability Paper No. 41).

Datastream. (1986–2015). *S&P 500 Composite Index; Spot Prices*. Thomson Reuters Datastream.

Datastream. (1986–2015). *S&P 500 index futures; 3-Month, 6-Month and 9-Month contracts*. Thomson Reuters Datastream.

Du, W., Tepper, A., & Verdelhan, A. (2018). Deviations from covered interest rate parity. *The Journal of Finance*, *73*(3), 915–957.

Duffie, D. (2010). Presidential address: Asset price dynamics with slow-moving capital. *The Journal of Finance*, *65*(4), 1237–1267.

Dwyer, G. P., Locke, P., & Yu, W. (1996). Index arbitrage and nonlinear dynamics between the S&P 500 futures and cash. *The Review of Financial Studies*, *9*(1), 301–332.

Federal Reserve. (1986–2015). *LIBOR*. Federal Reserve Economic Data Library.

Federal Reserve. (1986–2015). *Treasury Bill yields*. Federal Reserve Economic Data Library.

Fontaine, J.-S., & Garcia, R. (2011a). Bond liquidity premia. *The Review of Financial Studies*, *25*(4), 1207–1254.

Fontaine, J., & Garcia, R. (2011b). *The treasury security-based funding liquidity*. The website of Jean-Sebastien Fontaine.

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, *111*(1), 1–25.

Gallagher, L. A., & Taylor, M. P. (2001). Risky arbitrage, limits of arbitrage, and nonlinear adjustment in the dividend–price ratio. *Economic Inquiry*, *39*(4), 524–536.

Garleanu, N., & Pedersen, L. H. (2011). Margin-based asset pricing and deviations from the law of one price. *The Review of Financial Studies*, *24*(6), 1980–2022.

Gromb, D., & Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics*, *66*(2–3), 361–407.

Gromb, D., & Vayanos, D. (2010). Limits of arbitrage. *Annual Review of Financial Economics*, *2*, 251–275.

Gyntelberg, J., Hoerdahl, P., Ters, K., & Urban, J. (2017). *Arbitrage costs and the persistent non-zero CDS-bond basis: Evidence from intraday euro area sovereign debt markets* (BIS Working Paper No. 631).

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384.

Karnaukh, N., Ranaldo, A., & Söderlind, P. (2015). Understanding FX liquidity. *The Review of Financial Studies*, *28*(11), 3073–3108.

Kondo, J. E., & Papanikolaou, D. (2015). Financial relationships and the limits to arbitrage. *Review of Finance*, *19*(6), 2095–2138.

Martens, M., Kofman, P., & Vorst, T. C. (1998). A threshold error-correction model for intraday futures and index returns. *Journal of Applied Econometrics*, *13*(3), 245–263.

Mitchell, M., Pedersen, L. H., & Pulvino, T. (2007). Slow moving capital. *American Economic Review*, *97*(2), 215–220.

Mitchell, M., & Pulvino, T. (2012). Arbitrage crashes and the speed of capital. *Journal of Financial Economics*, *104*(3), 469–490.

Nagel, S. (2012). Evaporating liquidity. *The Review of Financial Studies*, *25*(7), 2005–2039.

Roll, R., Schwartz, E., & Subrahmanyam, A. (2007). Liquidity and the law of one price: The case of the futures-cash basis. *The Journal of Finance*, *62*(5), 2201–2234.

Schuster, P., & Uhrig-Homburg, M. (2015). Limits to arbitrage and the term structure of bond illiquidity premiums. *The Journal of Banking and Finance*, *57*, 143–159.

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, *52*(1), 35–55.

Stein, J. C. (2009). Presidential address: Sophisticated investors and market efficiency. *The Journal of Finance*, *64*(4), 1517–1548.

Tao, J., & Green, C. J. (2013). Transactions costs, index arbitrage and non-linear dynamics between ftse100 spot and futures: A threshold cointegration analysis. *International Journal of Finance & Economics*, *18*(2), 175–187.

Theissen, E. (2012). Price discovery in spot and futures markets: A reconsideration. *The European Journal of Finance*, *18*(10), 969–987.

Tse, Y. (2001). Index arbitrage with heterogeneous investors: A smooth transition error correction analysis. *The Journal of Banking and Finance*, *25*(10), 1829–1855.

van Binsbergen, J. H., Diamond, W. F., & Grotteria, M. (2019). *Risk-free interest rates* (Tech. Rep.). National Bureau of Economic Research.

How to cite this article: Chen, J., Cai, C. X., Faff, R., & Shin, Y. (2022). Nonlinear limits to arbitrage. *Journal of Futures Markets*, 1–30. https://doi.org/10.1002/fut.22320

APPENDIX A: PROOFS

In this appendix we provide the parametric assumption on the sensitivity parameter α to analyze the model in Section 2 and the proofs to Propositions 1–3 derived in Section 3.

Assumption A.1. The sensitivity parameter α is moderate, such that

$$\alpha < min\{\alpha^+, \alpha^*\},$$

where

$$\alpha^+ = \frac{V - (1-q)(S_{2b} - F_1)}{2(1-q)F_1},$$
$$\alpha^* = \frac{V - S_1 + F_1}{S_{2b} - S_1 + F_1}.$$

We impose two restrictions on α. First we note that the assumption of $\alpha < \alpha^+$ guarantees that α is not too extreme that arbitrageurs will lose their ability to bear against mispricing even under partial investment strategy, such that $\frac{\partial \kappa}{\partial S_1} < 0$. We will derive this parametric assumption under the proof of Proposition 1.

Second the assumption of $\alpha < \alpha^*$ is referred to as the stability condition, such that arbitrageurs cannot default in the bad state even when the full investment strategy is adopted. Recall F_{2b} from Equation (1) as

$$F_{2b} = F_1\left[1 + \alpha\left(\frac{P_{2b}}{P_1} - 1\right)\right].$$

By substituting P_{2b} and P_1 in Equations (2) and (3), we have

$$F_{2b} = F_1 - \alpha F_1\left[\frac{S_{2b} - S_1}{V - S_1 - (\alpha - 1)F_1}\right]. \tag{A1}$$

There are two paths to derive the condition for $F_{2b} > 0$. First, one can start by letting $V - S_1 - (\alpha - 1)F_1 < 0$, which can be simplified as

$$\alpha > \frac{V - S_1 + F_1}{F_1}.$$

This condition implies that $F_{2b} > F_1 > 0$ and $P_{2b} > P_1$. Note that in this scenario, the sensitivity α must be extremely large. Consider the numerical example from SV. Given that $V = 1$, $S_1 = 0.3$, $S_{2b} = 0.4$, $F_1 = 0.2$, then $\alpha > 4.5$. It implies that fund investors will more than quadruple the profits or losses that arbitrageurs achieved in the last period, which is rare in the real-world hedge fund industries.

Second, we allow $V - S_1 - (\alpha - 1)F_1 > 0$ and $F_{2b} > 0$, and derive the following condition as

$$\alpha < \alpha^* = \frac{V - S_1 + F_1}{S_{2b} - S_1 + F_1}. \tag{A2}$$

It suggests that $0 < F_{2b} < F_1$ and $P_{2b} < P_1$. Given that $V = 1$, $S_1 = 0.3$, $S_{2b} = 0.4$, $F_1 = 0.2$, then the condition becomes $\alpha < 3$, which is easily satisfied. Hence to keep the analysis tractable and reasonable, we impose Equation (A2) as the stability condition. We note that this stability condition is more general than the condition, $\alpha < \frac{V - S_1 + F_1}{F_1}$ imposed in SV.

Consider a relatively large $q \geq \frac{V - S_{2b} + F_1}{2V - S_{2b} + F_1}$.[20] Then, we have $\alpha^+ \geq \alpha^*$, and Assumption 1 can be simplified as $\alpha < \alpha^*$. Otherwise, we have $\alpha^+ < \alpha^*$, and Assumption 1 becomes $\alpha < \alpha^+$. Therefore to keep the analysis reasonable, we avoid extremely large α for any q:

$$\alpha < min\{\alpha^+, \alpha^*\}.$$

Proof of Lemma 1. Forced liquidation implies that arbitrageurs hold fewer shares of the asset in period 2, that is, $F_{2b}/P_{2b} < \hat{\beta}_1 F_1/P_1$. A strategy β_1^{liq} that ensures the equality holds, that is, $F_{2b}/P_{2b} = \beta_1^{liq} F_1/P_1$, is the threshold point where arbitrageurs hold the same number of shares in both periods 1 and 2. For any optimal strategy, $\hat{\beta}_1 > \beta_1^{liq}$, we must have forced liquidation in period 2, such that $F_{2b}/P_{2b} < \hat{\beta}_1 F_1/P_1$.

Recall F_{2b} from Equation (A1), P_1 and P_2 from Equations (2) and (3), the equality $F_{2b}/P_{2b} = \beta_1^{liq} F_1/P_1$ implies that $\beta_1^{liq} = \frac{V - S_1}{V - S_1 + (\alpha - 1)(S_{2b} - S_1)}$. □

Proof of Lemma 2. We first show the impact of S_1 on $\hat{\beta}_1$ and P_1. Consider the partial investment equilibrium in Equation (5):

$$(1 - q)\left(\frac{V}{P_1} - 1\right) + q\left(\frac{P_{2b}}{P_1} - 1\right)\frac{V}{P_{2b}} = 0.$$

The optimal $0 < \hat{\beta}_1 < 1$ can be derived as

$$\hat{\beta}_1 = \frac{n_1 - n_3}{2a(1 - q)F_1}, \tag{A3}$$

where

$$\begin{aligned} n_1 &= V + (1 - q)(F_1 + \alpha S_1 - S_{2b}), \\ n_2 &= V + (1 - q)(F_1 - \alpha S_1 - S_{2b}), \\ n_3 &= \sqrt{(n_2)^2 + 4\alpha V q(1 - q)(S_{2b} - F_1)}. \end{aligned}$$

As $\hat{\beta}_1$ has a complex expression, we simplify it with n_1, n_2, and n_3 as functions of model parameters. Then, the partial derivative of $\hat{\beta}_1$ with respect to S_1 is derived as

$$\frac{\partial \hat{\beta}_1}{\partial S_1} = \frac{1}{2F_1}\left(1 - \frac{n_2}{n_3}\right) > 0. \tag{A4}$$

$\frac{\partial \hat{\beta}_1}{\partial S_1} > 0$ since $n_3 > n_2$, which can be easily seen because of $F_1 < S_1 < S_{2b}$. This implies the positive capital allocation effect.

The partial derivative of P_1 with respect to S_1 is derived as

$$\frac{\partial P_1}{\partial S_1} = \frac{\partial \hat{\beta}_1}{\partial S_1} F_1 - 1 = \frac{1}{2}\left(1 - \frac{n_2}{n_3}\right) - 1 < 0, \tag{A5}$$

which holds since $n_3 > n_2$.

[20]The threshold, $\frac{V - S_{2b} + F_1}{2V - S_{2b} + F_1}$, is derived under $\alpha^* = \alpha^+$.

Next, we show how S_1 will affect the period-2 price in the bad state, P_{2b}. Consider the partial investment equilibrium. We rewrite Equation (5) as

$$P_{2b} = \frac{qVP_1}{V-(1-q)P_1}.$$

Taking the first differentiation with respect to S_1, then

$$\frac{\partial P_{2b}}{\partial S_1} = qV\left[\frac{\frac{\partial P_1}{\partial S_1}(V-(1-q)P_1)+\frac{\partial P_1}{\partial S_1}(1-q)P_1}{(V-(1-q)P_1)^2}\right] = \frac{\frac{\partial P_1}{\partial S_1}qV^2}{(V-(1-q)P_1)^2}.$$

Clearly, the sign of $\frac{\partial P_{2b}}{\partial S_1}$ is the same as that of $\frac{\partial P_1}{\partial S_1}$, which means that $\frac{\partial P_{2b}}{\partial S_1} < 0$ under the partial investment strategy.

Next, consider the full investment equilibrium. We obtain the partial derivative of F_{2b} with respect to S_1 as

$$\frac{\partial F_{2b}}{\partial S_1} = aF_1\frac{V-S_{2b}+(1-\alpha)F_1}{(V-S_1-(\alpha-1)F_1)^2} > 0.$$

The inequality holds under the stability condition such that

$$V-S_{2b}+(1-\alpha)F_1 > (\alpha-1)(S_{2b}-S_1) > 0.$$

Therefore, it is easily seen that $\frac{\partial P_{2b}}{\partial S_1} > 0$ for $\hat{\beta}_1 = 1$. □

Proof of Proposition 1. Proof of Proposition 1 is straightforward but rather tedious.[21]

Consider the partial investment equilibrium. It is straightforward to derive the partial differentiation of $\kappa = \frac{\hat{\beta}_1 F_1}{S_1}$ with respect to S_1 by

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

where $\hat{\beta}_1$ and $\frac{\partial \hat{\beta}_1}{\partial S_1}$ are defined in Equations (A3) and (A4), respectively. The numerator reaches its minimum at the corner solution, when $\hat{\beta}_1 = 1$ and the equality in FOC holds. We next solve S_1 at the corner solution:

$$\hat{S}_1 = F_1 + \frac{Vq(S_{2b}-F_1)}{V-(1-q)(S_{2b}-F_1+\alpha F_1)}.$$

The condition for the numerator being positive is now derived as

$$\frac{\partial \hat{\beta}_1}{\partial S_1}\hat{S}_1 > 1,$$

which can be simplified as

[21]The main analytic results have been derived using MATLAB12. Upon request the codes will be available.

$$\alpha < \alpha^{+} = \frac{V - (1-q)(S_{2b} - F_1)}{2(1-q)F_1}.$$

It implies that for any $\alpha < \alpha^{+}$, the partial derivative $\frac{\partial \kappa}{\partial S_1}$ is positive even when partially invested arbitrageurs reaches the corner solution at $\hat{\beta}_1 = 1$ and $S_1 = \hat{S}_1$. Hence we have $\frac{\partial \kappa}{\partial S_1} > 0$ under partial investment strategy.

Furthermore, under the partial investment equilibrium, we rearrange the FOC in Equation (5), and express λ as

$$\lambda = q\frac{V - P_{2b}}{V - P_1} = q + P_{2b}\frac{1-q}{V}. \tag{A6}$$

According to Lemma 2, we have $\frac{\partial P_{2b}}{\partial S_1} < 0$. Therefore, it is easily seen from Equation (A6) that as S_1 rises, λ falls. This proves that $\frac{\partial \kappa}{\partial S_1} > 0$ and $\frac{\partial \lambda}{\partial S_1} < 0$ under partial investment equilibrium.

Consider the full investment equilibrium, $\hat{\beta}_1 = 1$. In this case, we can express κ as

$$\kappa = \frac{\hat{\beta}_1 F_1}{S_1} = \frac{F_1}{S_1}.$$

Then, it is easily seen that $\frac{\partial \kappa}{\partial S_1} < 0$, given F_1. Furthermore, we rewrite λ as

$$\lambda = q\frac{V - P_{2b}}{V - P_1} = q\frac{V - P_{2b}}{S_1 - F_1}.$$

From Lemma 2, we have $\frac{\partial P_{2b}}{\partial S_1} > 0$. Hence, $\frac{\partial \lambda}{\partial S_1} < 0$. □

Proof of Proposition 2. q^* is derived at the point where arbitrageurs are indifferent to strategies (i.e., $D_1 = F_1$ under the partial investment strategy), and is determined as a complex function of the parameter set, $\{V, S_1, S_{2b}, F_1, \alpha\}$:

$$q^* = \frac{(S_1 - F_1)(V + F_1 - S_{2b} - \alpha F_1)}{V(S_{2b} - S_1) + (S_1 - F_1)(V + F_1 - S_{2b} - \alpha F_1)} = \frac{1}{m+1}, \tag{A7}$$

where $m = V(S_{2b} - S_1)/(S_1 - F_1)(V + F_1 - S_{2b} - \alpha F_1)$.

First, it is easily seen that

$$S_1 - F_1 > 0, \quad S_{2b} - S_1 > 0, \quad V - S_{2b} + (1-\alpha)F_1 > 0, \tag{A8}$$

where we use the maintained assumption that $F_1 < S_1 < S_{2b}$, and the stability condition, $\alpha \le \alpha^*$. It is then trivial to show that $m > 0$ and $0 < q^* < 1$. As S_1 approaches to F_1, the numerator in q^* becomes zero and $q^* \approx 0$. As S_1 approaches to S_{2b}, m becomes zero, and $q^* \approx 1$.

Next, taking the partial derivation of m with respect to S_1, we have

$$\frac{\partial m}{\partial S_1} = \frac{V(F_1 - S_{2b})}{(S_1 - F_1)^2(V + F_1 - S_{2b} - \alpha F_1)} < 0.$$

Using the inequality in Equation (A8), it is easily seen that $\frac{\partial q^*}{\partial S_1} > 0$ and $\frac{\partial q^*}{\partial F_1} < 0$. □

Proof of Proposition 3. We start with κ_q, such that

$$\kappa_q = \sum_{j \in J} P(q = j)\kappa(q = j),$$

where J denotes the set of all real number within the unit interval and $P(q = j)$ is the probability of $q = j$. It is straightforward to derive the partial differentiation of κ_q with respect to S_1 by

$$\frac{\partial \kappa_q}{\partial S_1} = P(q = j)\frac{\partial \kappa(q = j)}{\partial S_1}.$$

In one extreme where $S_1 = F_1$, we note that $q^* = 0$ from Proposition 2 and the partial investment strategy is optimal for any q. Hence we must have $\frac{\partial \kappa(q=j)}{\partial S_1} > 0$ for any $q = j$ from Proposition 1, and $\frac{\partial \kappa_q}{\partial S_1} > 0$. In another extreme where $S_1 = S_{2b}$, we note that $q^* = 1$ from Proposition 2 and the full investment strategy is optimal. Hence we have $\frac{\partial \kappa(q=j)}{\partial S_1} < 0$ for any $q = j$ from Proposition 1, and $\frac{\partial \kappa_q}{\partial S_1} < 0$.

For $F_1 < S_1 < S_{2b}$, we can derive the threshold probability q^*, and rewrite the partial differentiation as

$$\frac{\partial \kappa_q}{\partial S_1} = \sum_{j \in J} P(q = j)\left[\frac{\partial \kappa(q = j | j \geq q^*)}{\partial S_1} + \frac{\partial \kappa(q = j | j < q^*)}{\partial S_1}\right].$$

It is easily seen that the first term in the bracket $\frac{\partial \kappa(q=j \mid j \geq q^*)}{\partial S_1} > 0$ since the partial investment strategy is optimal, and the second term in the bracket $\frac{\partial \kappa(q=j \mid j < q^*)}{\partial S_1} < 0$ since the full investment strategy is optimal. Recall that Proposition 2 states that q^* increases with S_1. For S_1 increases from F_1 to S_{2b}, $\frac{\partial \kappa_q}{\partial S_1}$ is positive at first since q^* is small and the positive $\frac{\partial \kappa(q=j \mid j \geq q^*)}{\partial S_1}$ is dominating. But it turns to negative as S_1 grows higher since q^* is large and the negative $\frac{\partial \kappa(q=j \mid j < q^*)}{\partial S_1}$ is dominating.

Next we turn to λ_q, such that

$$\lambda_q = \sum_{j \in J} P(q = j)\lambda(q = j),$$

and the partial differentiation of λ_q with respect to S_1, such that

$$\frac{\partial \lambda_q}{\partial S_1} = \sum_{j \in J} P(q = j)\frac{\partial \lambda(q = j)}{\partial S_1}.$$

Since Proposition 1 states that $\frac{\partial \lambda(q=j)}{\partial S_1} < 0$, then we must have $\frac{\partial \lambda_q}{\partial S_1} < 0$.

It is worth noticing that in the above proof of Proposition 3, it does not require the actual distribution of q. It implies that results in Proposition 3 hold for any distribution of q. □

APPENDIX B: SIZE OF MISPRICING ERROR OVER TIME

Figure B1 plots the moving average of mispricing error in absolute value over the whole sample period 1982–2015. It is easily seen that mispricing error is large and volatile in the early period before 1986: it fluctuates above 0.25, and can reach as high as 2.25 at extreme. At the time when index futures contracts were first introduced in 1982, the market was characterized by high transaction costs, a low number of participating arbitrageurs and low levels of available arbitrage capital. Therefore larger mispricing errors tended to occur during the early periods. Over time, as knowledge diffused and entry barriers and implementation costs dropped, mispricing became more stable after 1986 and comoved with major market events. It stayed below 0.25 for most of the sample period, only exceeding this level at the time of a few extreme market events, such as the 1987 market crash and the 2007–2009 global financial crisis.

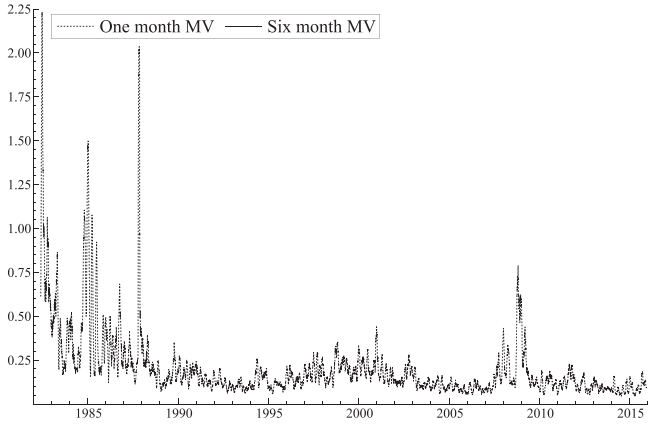

FIGURE B1 Plot of moving average of mispricing error, 1982–2015. The moving average of the absolute value of the mispricing error, $\hat{z}_t$, estimated from the long-run equation $f_t = \mu + \theta f_t^* + z_t$, where f_t and f_t^* are the spot and fundamental prices, respectively. The dotted line plots the 1-month moving-average mispricing error, while the solid line plots the 6-month moving-average mispricing error. MV, market value

APPENDIX C: OBSERVABLE VARIABLE DEFINITIONS

In this appendix, we introduce a number of variables as measures of funding and market liquidity.

1. The VIX index
   We collect the daily VIX data from the website of the CBOE.
2. Aggregate hedge fund flows, aggregate index fund flows, and aggregate mutual fund flows
   Ang et al. (2011) find that hedge fund leverage decreases more than the leverage of financial intermediaries. They also find that funding cost and fund return volatility can negatively predict fund leverage, while the market value (MV) of hedge funds positively predicts fund leverage. Empirically, the aggregate hedge fund flow (past 3-month flow) positively predicts the gross leverage and long-only leverage. We follow Ang et al. and construct the aggregate actively managed US mutual fund flows, aggregate passively managed US index fund flows, and aggregate US hedge fund flows. From the Morningstar Direct-defined universe of US Mutual Funds, we select the mutual funds with "Index funds" indicator "No" and "Oldest share class" indicator "Yes" (to remove identical funds with a different share class), and Global Category group "Equity" or "Allocation"; this yields 5152 funds as our mutual fund sample. Similarly, we select the index funds with "Index funds" indicator "No" and "Oldest share class" indicator "Yes" (a gain to remove identical funds), and Global Category group "Equity" or "Allocation"; this yields 291 funds as our index fund sample. From the Morningstar Global Hedge Fund universe, we select the funds with "Domicile" indicator "United States"; this yields 1451 funds as our hedge fund sample. All the return and total net asset data are downloaded on a monthly basis from January 1976 to December 2015. The monthly mutual fund/index fund/hedge fund flows are constructed as follows:

$$Flow_{i,t} = \frac{TNA_{i,t}}{TNA_{i,t-3}} - (1 + r_{i,t-2})(1 + r_{i,t-1})(1 + r_{i,t}),$$

$$AggreFlow_t = \sum_{i=1}^{k} Flow_{i,t},$$

   where $TNA_{i,t}$ is the total net assets of fund i in quarter t and $r_{i,t}$ is the total return of fund i in quarter t, obtained from the Morningstar Direct database.
3. Financial sector leverage We again follow Ang et al. (2011) to construct the financial sector leverage. The financial sector is defined to capture all US-based companies with Standard Industrial Classification codes between 6000 and 6299. Leverage for company i at quarter t is defined as

$$Leverage_{i,t} = \frac{\sum_{i \in sector} Asset_{i,t}}{\sum_{i \in sector} MV_{i,t}},$$

   where $Asset_{i,t}$ is the total assets of the company obtained from COMPUSTAT and $MV_{i,t}$ is the market value of the company obtained from CRSP.

4. Financial asset growth Adrian and Shin (2010) find that the growth in financial assets measures the increase of aggregate liquidity. Fast asset growth increases surplus capital as intermediaries seek to expand this capital and search for borrowers. Aggregate liquidity then rises as people are urged to borrow the money, though they may lack the capacity to repay it. Financial asset growth is constructed as follows:

$$AssetGrowth_{i,t} = \frac{Asset_{i,t}}{Asset_{i,t-1}} - 1,$$

where $Asset_{i,t}$ is the total financial assets of company i in quarter t obtained from COMPUSTAT.

5. Amihud (2002) illiquidity measure
We construct the Amihud (2002) illiquidity factor of stocks as follows:

$$Illiq_{i,m} = \left(\frac{1}{D_{i,m}}\right)\sum_{d=1}^{D_{i,m}} \frac{|R_{i,m,d}|}{VOLD_{i,m,d}},$$

where $D_{i,m}$ is the number of days of stock i available in month m, $R_{i,m,d}$ is the daily return of stock i on day d in month m, and $VOLD_{i,m,d}$ is the trading volume in dollars of stock i on day d in month m. We use securities that are traded on the New York Stock Exchange in the period from June 1986 and December 2015. Data is collected from the CRSP database.

6. Broker–dealer leverage Adrian et al. (2014) apply the broker–dealer leverage to measure the stochastic discount factor (SDF) when funding constraints are high. They find that worse funding conditions are related to deleveraging and high marginal value of wealth. The quarterly broker–dealer leverage factor of Adrian et al. (2014) is collected from Muir's webpage. This Broker–dealer leverage is constructed as follows:

$$Leverage_t^{BD} = \frac{Total\ Financial\ Assets_t^{BD}}{Total\ Financial\ Assets_t^{BD} - Total\ Liabilities_t^{BD}},$$

where $Total\ Financial\ Assets_t^{BD}$ is the aggregate quarterly total financial assets of security broker–dealers, and $Total\ Liabilities_t^{BD}$ is the aggregate quarterly total financial liabilities of security broker–dealers reported.

7. Treasury security-based funding liquidity
The treasury security-based funding liquidity (Fontaine & Garcia, 2011a) data are directly obtained from the website of Jean-Sebastien Fontaine.

8. TED spread
TED spread is constructed as follows:

$$TED_t = Yield_{EU,t} - Yield_{US,t},$$

where $Yield_{EU,t}$ is the yield of 3-month Eurodollar deposits (LIBOR), and $Yield_{US,t}$ is the yield of 3-month US T-bills. The TED spread is the 3-month US T-bill deposit yield subtracted from the 3-month Eurodollar deposit yield (LIBOR). Both LIBOR and T-bill yields are monthly data downloaded from the Federal Reserve Economic Data (FRED) library.

APPENDIX D: TESTS OF ROBUSTNESS

D.1 | Evidence with alternative measure of risk-free rate

We provide tests of robustness with an alternative measure of risk-free rate. Table D1 and Figure D1 duplicate our estimation of MS-GECM on the S&P 500 index spot-futures relation with the 3-month LIBOR rate as the risk-free interest rate. Similarly, the initial mispricing correction displays the inverse U-shape as mispricing widens, while noise momentum sharply declines in the extreme regime.

TABLE D1 Estimation of the Markov-switching generalized error correction model, 3-month LIBOR rate

	State 1		State 2		State 3	
	Estimate	***t* stat**	**Estimate**	***t* stat**	**Estimate**	***t* stat**
Panel A. Estimation results						
α	0.009***	4.80	0.003	0.57	−0.109**	2.30
δ	0.991***	319.0	1.016***	254.0	1.156***	50.6
γ	−0.003	−1.28	0.013***	4.06	0.098***	4.14
κ	−0.699***	−38.0	−0.883***	−43.8	−0.631***	−7.66
λ^*	0.262***	14.9	0.109***	5.90	0.123	1.56
Σ	0.112***	49.9	0.224***	34.9	0.953***	15.0
SOA	0.436***	19.9	0.773***	31.8	0.507***	4.68
$\lvert z_{t-1}\rvert$	0.101		0.187		0.685	
Log-likelihood		2504.23				
Panel B. Recovered coefficients						
λ	0.871***	9.77	0.935***	3.75	0.333	1.42

Note: This table reports the estimation of the Markov-switching GECM. The sample covers the daily series of the S&P 500 index and its 3-month-to-maturity futures contracts over the period June 4, 1986–December 3, 2015. There are a total of 7442 observations, of which 3973, 3279, and 190 fall into States 1, 2, and 3, respectively. Specifically, Panel A reports the estimation results for $\Delta f_t = \alpha_{R_j} + \kappa_{R_j}\hat{z}_{t-1} + \lambda^*_{R_j}\hat{z}_{t-2} + \delta_{R_j}\Delta f^*_t + \gamma_{R_j}\Delta f_{t-1} + \mu_{tR_j}$, where $\hat{z}_t$ is estimated from Equation (11), the risk-free rate is proxied as the 3-month LIBOR rate, $\{\alpha_{R_j}, \delta_{R_j}, \gamma_{R_j}, \kappa_{R_j}, \lambda^*_{R_j}\}$ are regime-dependent coefficients with the covariance of the residuals (Σ_{R_j}), taking different values across the three states. Panel B reports the recovered coefficients. Specifically, $\lambda_{R_j} = \lambda^*_{R_j}/(1+\kappa_{R_j})$. For nonlinear combinations of the coefficients, a delta method is applied to obtain the variance of the recovered coefficients and their differences. All *t* statistics are computed based on a numerical Hessian matrix, and ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

Abbreviations: GECM, generalized error correction model; SOA, overall speed of adjustment.

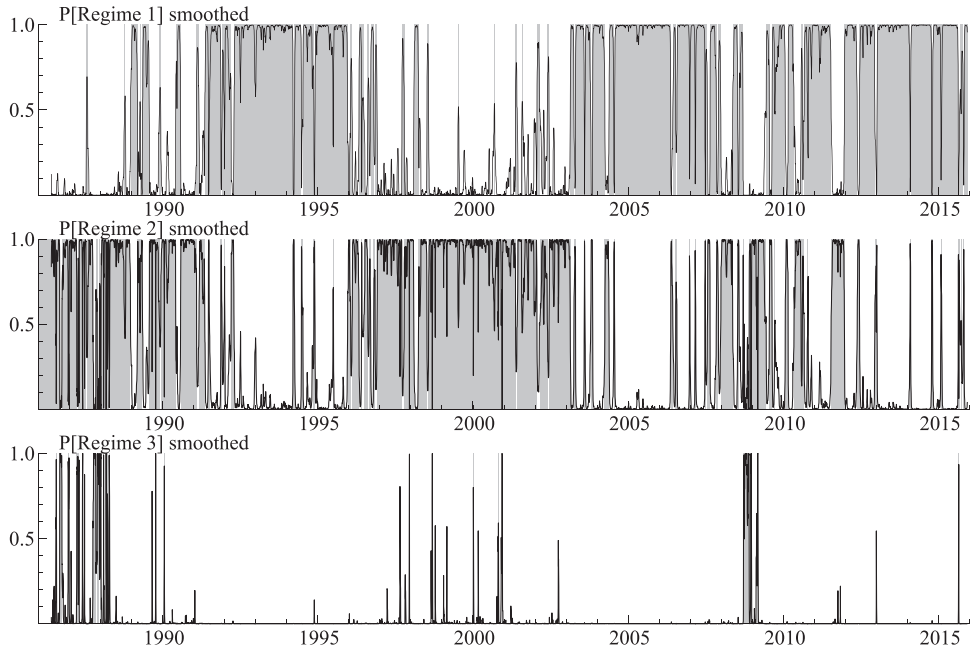

FIGURE D1 The smoothed regime probability, 3-month LIBOR rate. The smoothed regime probability of being in State 1, 2, or 3. It is resulted from the estimation of MS-GECM with 3-month-to-maturity futures contracts over 1986–2015, while using the 3-month LIBOR rate as the risk-free interest rate. State 1 consists of the years 1992–1995, 2003–2007, and 2012–2015, State 2 consists of the years 1986–1991, 1996–2002, 2009, and 2011, while State 3 is found in the years of 1987, 1998, 2001, and 2008. MS-GECM, Markov-switching generalized error correction model

TABLE D2 Estimation of the Markov-switching generalized error correction model, 3-month-to-maturity futures contracts, 1990–2015

	State 1		State 2		State 3	
	Estimate	***t* stat**	**Estimate**	***t* stat**	**Estimate**	***t* stat**
Panel A. Estimation results						
α	−0.010***	−3.95	0.021***	4.39	0.109**	2.30
δ	0.992***	301.0	1.012***	241.0	1.006***	73.0
γ	−0.005*	−1.74	0.017***	4.76	0.009	0.71
κ	−0.709***	−35.8	−0.802***	−37.6	−0.806***	−10.8
λ^*	0.230***	11.8	0.164***	7.87	0.098	1.25
Σ	0.109***	49.9	0.218***	34.9	0.567***	15.0
SOA	0.479***	18.5	0.637***	22.4	0.707***	6.56
$\lvert z_{t-1} \rvert$	0.099		0.188		0.498	
Log-likelihood		2582.75				
Panel B. Recovered coefficients						
λ	0.793***	13.1	0.827***	9.09	0.510***	5.75

Note: This table reports the estimation of the Markov-switching GECM. The sample covers the daily series of the S&P 500 index and its 3-month-to-maturity futures contracts over the period June 1, 1990–December 3, 2015. There are a total of 6430 observations, of which 3629, 2630, and 171 fall into States 1, 2, and 3, respectively. Specifically, Panel A reports the estimation results for $\Delta f_t = \alpha_{R_j} + \kappa_{R_j}\hat{z}_{t-1} + \lambda^*_{R_j}\hat{z}_{t-2} + \delta_{R_j}\Delta f^*_t + \gamma_{R_j}\Delta f_{t-1} + \mu_{tR_j}$, where $\hat{z}_t$ is estimated from Equation (11), $\{\alpha_{R_j}, \delta_{R_j}, \gamma_{R_j}, \kappa_{R_j}, \lambda^*_{R_j}\}$ are regime-dependent coefficients with the covariance of the residuals (Σ_{R_j}), taking different values across the three states. Panel B reports the recovered coefficients. Specifically, $\lambda_{R_j} = \lambda^*_{R_j}/(1+\kappa_{R_j})$. For nonlinear combinations of the coefficients, a delta method is applied to obtain the variance of the recovered coefficients and their differences. All *t* statistics are computed based on a numerical Hessian matrix, and ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

Abbreviations: GECM, generalized error correction model; SOA, overall speed of adjustment.

D.2 | Evidence with alternative sample periods

We provide tests of robustness with alternative sample periods and S&P 500 futures contracts. Table D2 duplicates our estimation of MS-GECM on the S&P 500 index spot-futures relation with a shorter sample period. We drop the periods with the stock market crash in 1987, and start from June 1990. Results in States 1 and 2 are similar to those in Table 2, with κ increasing by 10% while λ remaining stable as mispricing increases from States 1 to 2. It implies that the positive capital allocation effect dominates the negative funding constraint effect from States 1 to 2. Arbitrage activity becomes more aggressive against larger mispricing from States 1 to 2. κ in State 3 (80%) is rather similar to that in State 2 despite the significant growth in mispricing error (from 0.188 in State 2 to 0.498 in State 3). It is hard to tell whether the binding funding constraints deter the initial mispricing correction or the marginal impact of mispricing is compressed in the presence of high correction and large mispricing. The subsequent noise momentum reveals vital information. λ witnesses a sharp decline from 80% to 50%, which confirms the dominance of the negative funding constraint effect in the extreme regime (e.g., State 3 in Table 2). The interplay between κ and λ against mispricing is consistent with the nonlinear limits to arbitrage.

D.3 | Evidence with alternative futures contracts

We also provide evidence with the S&P 500 futures contracts that have 6 and 9 months to maturity, which are also traded actively. Comparing to the 3-month-to-maturity contracts, the 6- and the 9-month-to-maturity contracts will roll over on the third Friday of every quarter (March, June, September, and December) into the successive contracts. Tables D3 and D4, and Figure D2 duplicate the estimation of MS-GECM using contracts that have 6 and 9 months to maturity over 1990–2015. For both 6- and 9-month contracts, the interplay between κ and λ against mispricing is similar to the results with the 3-month contracts. κ increases significantly and λ declines slightly from States 1 to 2 with larger mispricing, resulting in a faster speed of adjustment. From States 2 to 3 where mispricing grows to an extreme level, arbitrage activities are rather deterred, as κ stops climbing and λ sharply drops. These results again highlight the dominance of the funding constraint effect in limiting arbitrage activity during the extreme State 3.

TABLE D3 Estimation of the Markov-switching generalized error correction model, 6-month-to-maturity futures contracts, 1990–2015

	State 1		State 2		State 3	
	Estimate	***t* stat**	**Estimate**	***t* stat**	**Estimate**	***t* stat**
Panel A. Estimation results						
α	−0.014***	−4.82	0.021***	3.98	0.153***	3.22
δ	0.988***	261.0	1.013***	231.0	1.003***	75.0
γ	−0.005	−1.64	0.013***	3.49	0.010	0.79
κ	−0.671***	−32.2	−0.739***	−34.2	−0.738***	−11.4
λ^*	0.288***	14.0	0.214***	10.0	0.184***	2.70
Σ	0.108***	46.8	0.227***	32.1	0.561***	8.22
SOA	0.383***	14.2	0.525***	18.4	0.554***	5.95
$\lvert z_{t-1}\rvert$	0.109		0.206		0.527	
Log-likelihood		2157.24				
Panel B. Recovered coefficients						
λ	0.875***	15.7	0.820***	11.6	0.705***	9.28

Note: This table reports the estimation of the Markov-switching GECM. The sample covers the daily series of the S&P 500 index and its 6-month-to-maturity futures contracts over the period June 1, 1990–December 3, 2015. There are a total of 6430 observations, of which 3500, 2698, and 232 fall into States 1, 2, and 3, respectively. Specifically, Panel A reports the estimation results for $\Delta f_t = \alpha_{R_j} + \kappa_{R_j}\hat{z}_{t-1} + \lambda^*_{R_j}\hat{z}_{t-2} + \delta_{R_j}\Delta f_t^* + \gamma_{R_j}\Delta f_{t-1} + \mu_{tR_j}$, where $\hat{z}_t$ is estimated from Equation (11), $\{\alpha_{R_j}, \delta_{R_j}, \gamma_{R_j}, \kappa_{R_j}, \lambda^*_{R_j}\}$ are regime-dependent coefficients with the covariance of the residuals (Σ_{R_j}), taking different values across the three states. Panel B reports the recovered coefficients. Specifically, $\lambda_{R_j} = \lambda^*_{R_j}/(1+\kappa_{R_j})$. For nonlinear combinations of the coefficients, a delta method is applied to obtain the variance of the recovered coefficients and their differences. All *t* statistics are computed based on a numerical Hessian matrix, and ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

Abbreviations: GECM, generalized error correction model; SOA, overall speed of adjustment.

TABLE D4 Estimation of the Markov-switching generalized error correction model, 9-month-to-maturity futures contracts, 1990–2015

	State 1		State 2		State 3	
	Estimate	***t* stat**	**Estimate**	***t* stat**	**Estimate**	***t* stat**
Panel A. Estimation results						
α	−0.007***	−3.32	0.016***	3.06	0.278***	3.88
δ	0.984***	262.0	1.015***	232.0	1.006***	65.0
γ	−0.006**	−2.03	0.012***	3.18	0.009	0.65
κ	−0.552***	−30.4	−0.639***	−29.6	−0.642***	−9.67
λ^*	0.395***	22.0	0.285***	13.8	0.242***	3.52
Σ	0.112***	45.0	0.245***	39.9	0.652***	17.9
SOA	0.157***	7.17	0.353***	13.0	0.399***	4.23
$\lvert z_{t-1}\rvert$	0.157		0.248		0.839	
Log-likelihood		1805.78				
Panel B. Recovered coefficients						
λ	0.883***	20.9	0.792***	15.0	0.677***	11.9

Note: This table reports the estimation of the Markov-switching GECM. The sample covers the daily series of the S&P 500 index and its 9-month-to-maturity futures contracts over the period June 1, 1990–December 3, 2015. There are a total of 6430 observations, of which 3541, 2662, and 227 fall into States 1, 2, and 3, respectively. Specifically, Panel A reports the estimation results for $\Delta f_t = \alpha_{R_j} + \kappa_{R_j}\hat{z}_{t-1} + \lambda^*_{R_j}\hat{z}_{t-2} + \delta_{R_j}\Delta f_t^* + \gamma_{R_j}\Delta f_{t-1} + \mu_{tR_j}$, where $\hat{z}_t$ is estimated from Equation (11), $\{\alpha_{R_j}, \delta_{R_j}, \gamma_{R_j}, \kappa_{R_j}, \lambda^*_{R_j}\}$ are regime-dependent coefficients with the covariance of the residuals (Σ_{R_j}), taking different values across the three states. Panel B reports the recovered coefficients. Specifically, $\lambda_{R_j} = \lambda^*_{R_j}/(1+\kappa_{R_j})$. For nonlinear combinations of the coefficients, a delta method is applied to obtain the variance of the recovered coefficients and their differences. All *t* statistics are computed based on a numerical Hessian matrix, and ***, **, and * indicate significance at 1%, 5%, and 10% levels, respectively.

Abbreviations: GECM, generalized error correction model; SOA, overall speed of adjustment.

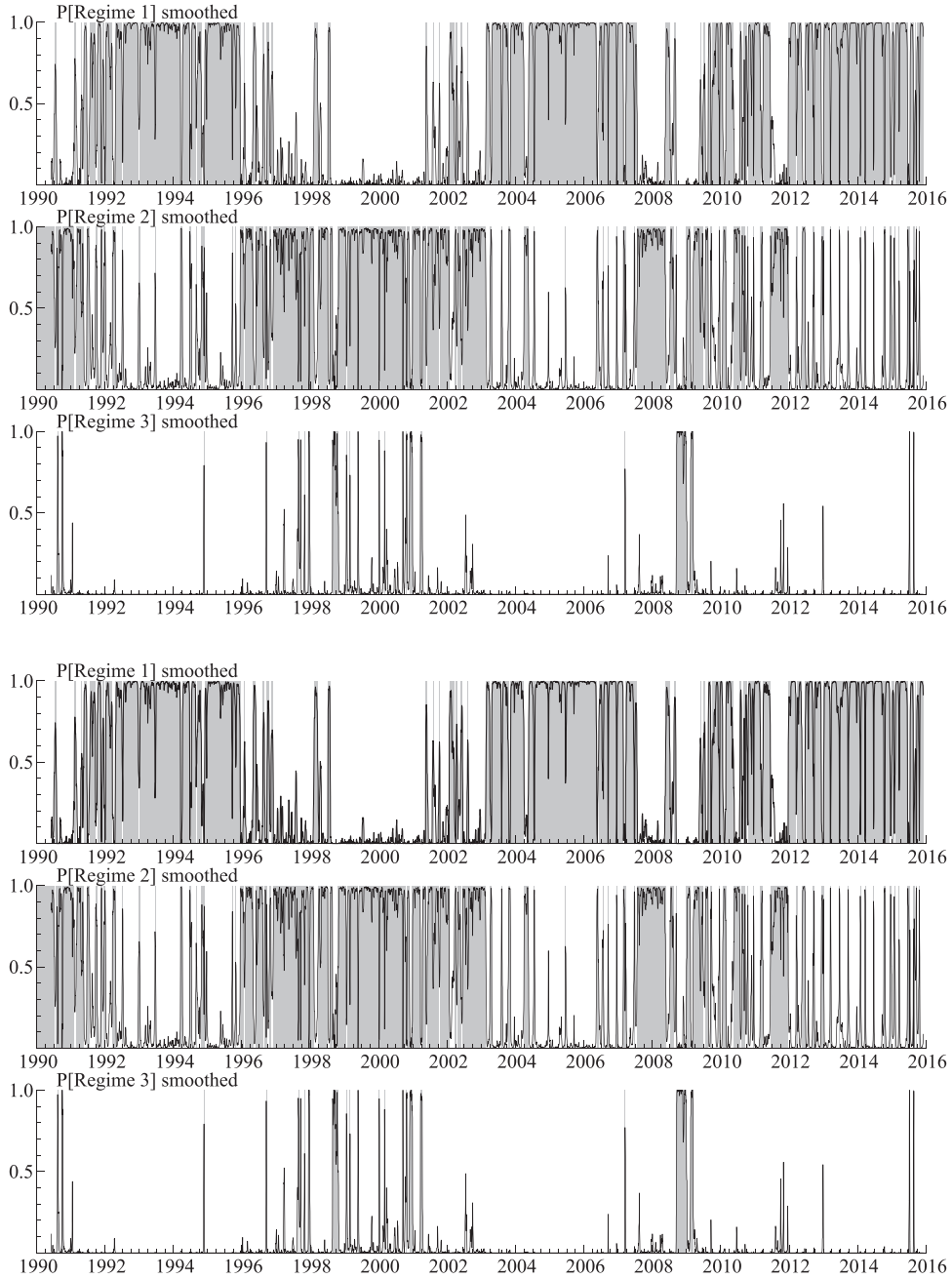

FIGURE D2 The smoothed regime probability, 1990–2015. All three graphs in the figure plot the smoothed regime probability of being in State 1, 2, or 3, estimated with futures contracts of different times to maturity. They are resulted from the estimation of MS-GECM over 1990–2015, with the 3-month-to-maturity futures contracts (top panel), the 6-month-to-maturity futures contracts (middle panel), and the 9-month-to-maturity futures contracts (bottom panel). MS-GECM, Markov-switching generalized error correction model

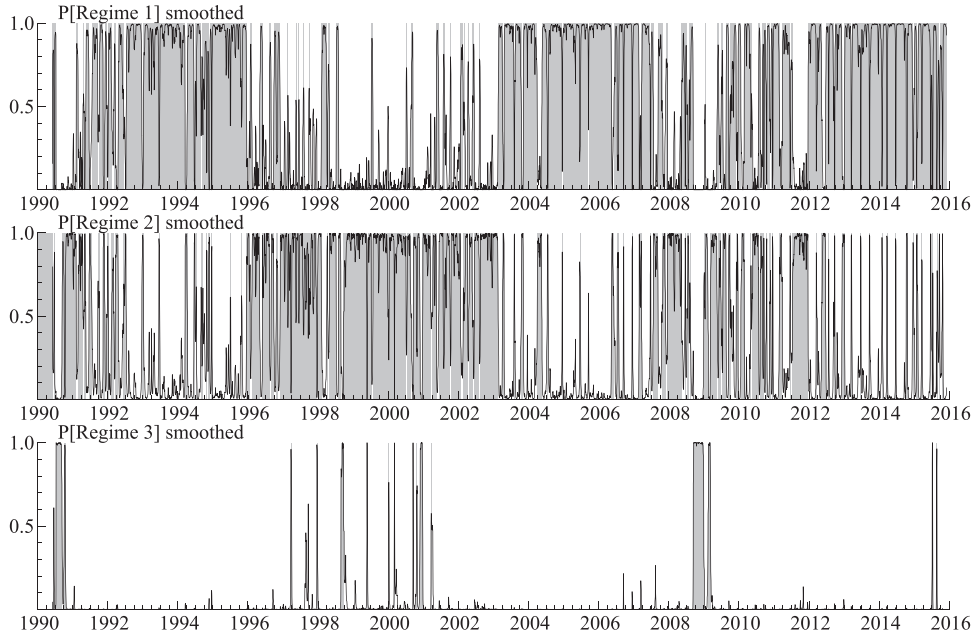

FIGURE D2 (Continued)