Financial Conduct Authority

68

Occasional Paper

May 2025

ETF (Mis)pricing

Andrei Kirilenko, Wladimir Kraus, Oliver Linton and Mingmei Xiao

FCA occasional papers in financial regulation

The FCA occasional papers

The FCA is committed to encouraging debate on all aspects of financial regulation and to creating rigorous evidence to support its decision-making. To facilitate this, we publish a series of Occasional Papers, extending across economics and other disciplines.

The main factor in accepting papers is that they should make substantial contributions to knowledge and understanding of financial regulation. If you want to contribute to this series or comment on these papers, please contact Kieran Keohane (kieran.keohane@fca.org.uk) or David Stallibrass (david.stallibrass@fca.org.uk)

Disclaimer

Occasional Papers contribute to the work of the FCA by providing rigorous research results and stimulating debate. While they may not necessarily represent the position of the FCA, they are one source of evidence that the FCA may use while discharging its functions and to inform its views. The FCA endeavours to ensure that research outputs are correct, through checks including independent referee reports, but the nature of such research and choice of research methods is a matter for the authors using their expert judgement. To the extent that Occasional Papers contain any errors or omissions, they should be attributed to the individual authors, rather than to the FCA.

Authors

Andrei Kirilenko is with University of Cambridge.

Wladimir Kraus is with the FCA.

Oliver Linton is with University of Cambridge.

Mingmei Xiao is with University of Cambridge and the FCA.

Acknowledgements

We would like to thank Karen Croxson, Paul Flanagan, Mark Northwood, Lucas Pedace, and Gian Giacomo Valentini for their help in preparing this study.

All our publications are available to download from www.fca.org.uk. If you would like to receive this paper in an alternative format, please call 020 7066 9644 or email publications_graphics @fca.org.uk or write to Editorial and Digital Department, Financial Conduct Authority, 12 Endeavour Square, London E20 1JN.

Summary

Authorised Participants (APs), primarily market makers, possess the right to create and redeem Exchange Traded Funds (ETF) shares based on market demand. The important role they play in facilitating liquidity provision and eliminating ETF mispricing makes their behaviour crucial to the well-functioning of the ETF market. Using a novel regulatory dataset that covers the primary and secondary market transactions of 128 ETFs from 2018 to 2022, we identify a connection between mispricing (the difference between ETF prices and the Net Asset Value (NAV) of their underlying baskets) and AP's inventory. We found that the skill of specialized traders (APs) in managing inventory and the overall demand for an ETF are important reasons why its price might temporarily be "wrong." Our model predicted this, and our real-world data backs it up, showing these factors add explanatory power on top of standard economic or fundamental influences. Further, our model is helpful for understanding the incentive structure of APs’ market making and arbitraging, as well as the mechanisms behind the significant mispricing observed in March 2020 across various ETF classes.

1 Introduction

Exchange Traded Funds (ETFs) are investment companies that pool investor capital to invest in underlying assets such as stocks or bonds. These funds are designed, issued, and managed by regulated financial institutions known as ETF sponsors. To ensure that ETF shares can be effectively created or redeemed to meet investor demand, sponsors contract with specialised financial entities called Authorised Participants (APs).

APs facilitate the creation and redemption process by engaging in primary market transactions, typically in-kind exchanges of the underlying assets, directly with the sponsor. The APs then trade these ETF shares with investors in the secondary market. A crucial aspect of their secondary market activity involves acting as market makers. This market-making role is intended to provide liquidity and help keep the ETF share price close to its Net Asset Value (NAV). However, deviations between ETF price and NAV -- termed ETF mispricing -- can and do occur.

Through their combined primary market creation/redemption activities and their secondary market making, APs accumulate net inventory of ETF shares. Maintaining this inventory is costly due to balance sheet constraints, leading APs to impose individual limits on their holdings. As inventory approaches these limits, particularly in stressed market conditions, APs' management actions such as offering shares at discounted prices -- can exacerbate mispricing for investors.

The importance of understanding these dynamics was starkly highlighted during the COVID-19 pandemic, when large mispricings in bond ETFs emerged as a significant financial stability concern. The International Monetary Fund (October 2022) reported that during the March 2020 market stress, the difference between NAV and price on bond ETFs dramatically increased, reaching over 5% across all bond ETFs and substantially more for specific categories like high-yield and investment-grade bond ETFs. This period of intense market dislocation prompted extraordinary central bank intervention, including the U.S. Federal Reserve's establishment of the Secondary Market Corporate Credit Facility (SMCCF), which purchased corporate bonds and bond ETFs to support market liquidity, underscoring the systemic importance of ETF market functioning.

To illustrate the main points, Figure 1 presents ETF Mispricing (ETF price minus NAV) for iShares Core MSCI World UCITS ETF. This ETF makes investments in a wide range of global companies in 23 developed countries. It covers 85 percent of listed equities in each country. It was issued in 2009 and its market cap [in 2024] is over GBP 50 billion. This liquid ETF experienced a negative 5 percent mispricing on 13 March 2020, followed by a positive 3.4 mispricing on 16 March 2020, and then elevated mispricing for months afterwards. Mispricing gradually subsided by the end of 2020 only to reappear in 2022 and remain elevated until the end of our sample.

While the existing literature has explored ETF mispricing and AP arbitrage (See Falato et al. (2021), Aramonte and Avalos (2020)), a persistent challenge has been the direct observation of APs' inventory positions. Many studies rely on lower-frequency data (e.g., annual N-CEN filings used by Gorbatikov and Sikorskaya (2022) and Raddatz (2021)) or aggregated holdings that do not distinguish individual AP inventory (e.g., ETF Global data used by Shim and Todorov (2023) and Koont et al. (2022)) making it difficult to precisely link APs inventory management to mispricing.

This paper presents both theoretical and empirical analysis of ETF mispricing due to APs inventory management activities powered by a novel dataset that combines both primary and secondary market transaction data. Primary market transactions data, i.e. exact creation and redemption of ETF shares was obtained directly from two major ETF sponsors. Secondary market transaction data comes from the UK Financial Conduct Authority (FCA) MiFID II regulatory database. This unique data combination allows us, for the first time, to precisely calculate daily inventory levels for individual APs across a sample of 128 ETFs (50 equity and 78 bond ETFs) from January 2018 to October 2022. Our work extends descriptive studies which used similar data sources (Aquilina et al. (2020) and Aquilina et al. (2021)) by covering the volatile COVID-19 pandemic period and, crucially, by developing and validating a dynamic theoretical model of ETF mispricing.

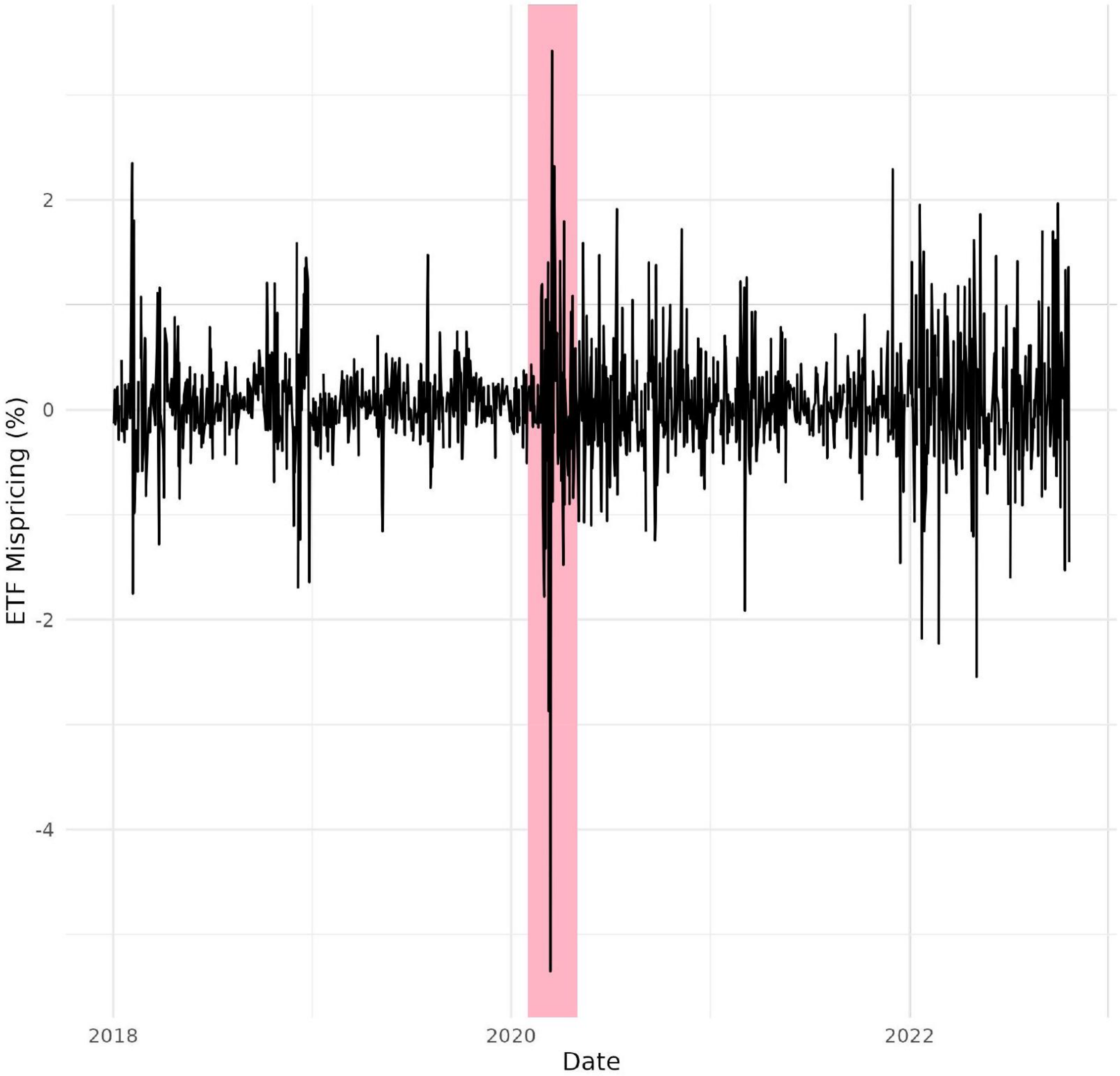
Our central research question is: How does APs' inventory activity influence ETF mispricing, as manifested under varying market conditions and across different ETF asset classes? To investigate this, we first develop a theoretical model of representative AP's optimal inventory management. The model considers costs associated with market making, arbitrage, and inventory holding, initially in a static setting and then extended to a dynamic setting. The dynamic model incorporates informed and noise traders, as well as APs learning about the ETF's fundamental value. We then empirically test the model's predictions using our unique granular dataset, which involves constructing key unobservable variables such as ETF fundamental values, APs' time-varying optimal inventory levels, and daily unexpected order imbalances.

We find that AP inventory levels significantly impact ETF mispricing, primarily by influencing ETF prices rather than their NAVs; for instance, excess AP inventory tends to widen discounts. These model-implied dynamics appear to be most representative for APs acting as high-frequency traders or dedicated market makers. Our empirical results also show significant heterogeneity: the effect of unexpected order imbalances is stronger for fixed-income ETFs, while the inventory effect is more pronounced for equity ETFs. Furthermore, our evidence suggests that APs sometimes deviate from purely arbitrage-driven inventory management, potentially taking on directional positions that reflect longer inventory management timeframes or other incentives, contrasting with views that APs consistently and immediately attempt to correct all mispricing (cf. Laipply and Madhavan, 2020))

Overall, our results demonstrate that APs' dual roles as market makers and arbitrageurs, coupled with their inventory constraints and beliefs about fundamental value, are crucial determinants of ETF price dynamics and market quality. These factors can significantly influence their ability to address ETF mispricing, particularly during periods of significant market dislocation.

This paper is organized as follows. Section 2 introduces the institutional background of ETFs. Section 3 illustrates our models of APs' ETF arbitrage. In Section 4, we describe our data sources and present the summary statistics of our dataset. Section 5 empirically constructs useful measures we use in the empirical section and Section 6 empirically examines the model-implied dynamics. Section 7 concludes.

Figure 1: iShares Core MSCI World UCITS ETF Premium During Covid



This figure demonstrates the ETF premium defined as 100(price-NAV)/NAV for iShares Core MSCI World UCITS ETF for our sample period: between Jan 22018 and Oct 22, 2022. This ETF has broad exposure to a wide range of global companies in 23 developed countries. It covers 85% of listed equities in each country. It was issued in 2009 and its current market cap is over GBP 50 billion. The red shaded area highlights the period from 01 Feb 2020 to 01 May 2020 where we observe a -5% mispricing on 13 March 2020, followed by a 3.4% mispricing on 16 March 2020. We provide a focus on that period in Figure 10.*

2 Institutional Background

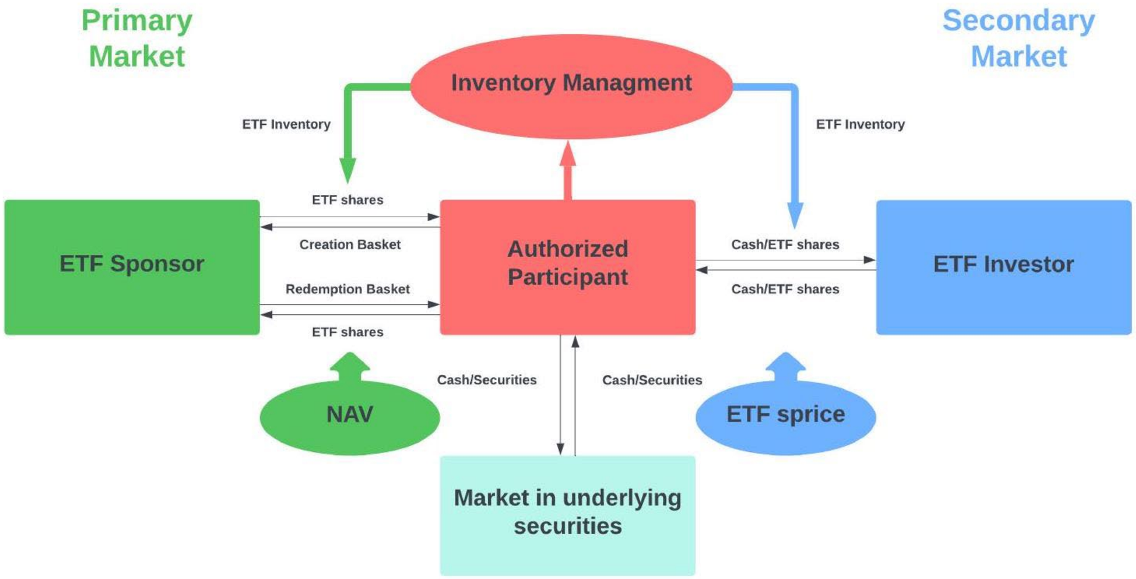
Exchange-Traded Funds (ETFs) are investment vehicles that combine features of both mutual funds and individual stocks, providing a unique blend of diversification, liquidity, and flexibility. Like mutual funds, ETFs can be created and redeemed in the primary market. Unlike mutual funds, however, ETFs are listed and can be traded on the stock exchange, similar to individual stocks. Investors can buy and sell ETF shares throughout the trading day at market prices, which may fluctuate based on supply and demand. This intraday trading feature provides flexibility and liquidity, allowing investors to react quickly to market conditions.

ETFs typically have lower expense ratios compared to mutual funds. The passive management strategy, combined with the creation and redemption mechanism, minimizes trading costs, and reduces the need for active management. Additionally, the in-kind creation and redemption process helps mitigate capital gains distributions, making ETFs more tax-efficient for investors. The feature of in-kind creation/redemption also helps mitigate against runs on the fund (commonly seen for mutual funds) during stress periods as the cost of redemption and liquidation is borne by the redeeming investor instead of the remaining investors in the fund. This removes the first mover advantage and makes ETFs more resilient during stress periods Falato et al. (2021).

ETFs are created and redeemed through a process involving authorised participants (APs), which are typically large financial institutions including banks, broker dealers, and principal trading firms (Aquilina et al. (2021)). For a physical ETF[1] when an AP wants to create new ETF shares, it delivers a basket of the underlying securities to the ETF provider in exchange for an equivalent number of ETF shares. This basket usually mirrors the composition of the ETF's target index. The AP can then sell these ETF shares on the open market. Conversely, when an AP wants to redeem ETF shares, it returns the ETF shares to the provider in exchange for the underlying securities. This mechanism helps maintain the ETF's price close to its net asset value (NAV). This process is clearly illustrated by Figure 2.

[1] All our sample ETFs are physical. There are also synthetic ETFs, which are a type of exchange-traded fund that seeks to replicate the performance of a benchmark index using derivatives rather than holding the actual underlying assets. Instead of directly purchasing the securities that constitute the index, synthetic ETFs enter into swap agreements with counterparty financial institutions. These swaps allow the ETF to receive the return of the index in exchange for a fee. See Investopedia article on Synthetic ETFs for more information.

Figure 2: ETF Trading Process



Adapted from Aquilina et al. (2020).

The ETF issuers usually disclose their holdings at the start of each trading day and the conversion of the basket of underlying into ETF shares (or the reverse) happens at the end of the trading day for "in-kind" creation/redemption. "In-cash" creation/redemption happens the at the beginning of the next trading day, and this is when the ETF issuer accepts cash value of the underlying basket instead of the physical underlying, in exchange for the ETF shares. For some ETFs, APs incur a cost for creation/redemption, but such costs are negligible for our sample ETFs.

It's important to note that APs enter into a legal contract with the issuer to obtain the right, instead of the obligation to create/redeem ETFs. Thus, APs are usually also ETF market makers and sometimes underlying broker/dealers who have good understanding of the current market demand. However, because of dual/multiple roles, there can be clashes of incentives due to cost and risk management in these roles.

For more detailed information on ETFs in general, please refer to Lettau and Madhavan (2018) and Aquilina et al. (2020).

3 Model

3.1 A static model of AP's ETF trades

We present a stylized model of a representative AP's optimal inventory management decisions. These decisions involve two assets in the model economy - one ETF and one underlying asset, in which one share of the ETF is issued to invest in one share of the underlying asset. We assume that the AP makes markets in the ETF and arbitrages between the ETF and the underlying asset. Without the loss of generality, we set the initial value of the underlying asset (NAV) to be equal to the price of the ETF ($NAV = P_0$), i.e., initial mispricing is normalised to zero. Furthermore, without the loss of generality, we normalize the AP's initial inventory in both markets to zero.

In addition to the AP, there is an uninformed ETF investor who needs to exogenously trades X shares of the ETF in the perfectly competitive secondary market. The AP, who acts as a market maker, does not initially know what X is going to be, but does know that X is drawn from a distribution with mean zero (i.e., the investor is equally likely to buy or sell ETF shares) and a known exogenous variance $\sigma_X^2 > 0$. The AP optimally sets their inventory management decisions in the initial period before X is realised.

When X is realised, the AP sells X shares of the ETF at the perfectly competitive price, P_e per share, which the AP sets to make zero profits. To sell X shares of the ETF, the AP chooses a proportion $0 \leq \gamma \leq 1$ of the ETF demand to satisfy with a shorting position. The AP creates $(1-\gamma)X$ shares of the ETF with the ETF sponsor/issuer in an in-kind exchange with the underlying asset that the AP bought from the underlying market.

We assume that the AP's market making, and arbitrage actions are costly because (i) by choosing to hold a fraction of ETF shares on her balance sheet as inventory, the AP incurs a financing cost, and (ii) by purchasing the underlying asset for creation, the AP impacts the price and, hence, the NAV in the market for the underlying.

We assume that inventory holding cost increases quadratically in the number of held shares, i.e., $\frac{\lambda\sigma_X^2}{2}(\gamma X)^2$ with an exogenous $0 < \lambda < \bar{\lambda}$. The quadratic cost assumption is consistent with the non-linear premium associated with the funding cost of the AP's balance sheet as argued by Shen (2002). We assume that the upper bound on λ is such that the holding costs are not prohibitively high to result in AP's creating all X shares of the ETF. Intuitively, inventory holding cost reflects both funding conditions for individual APs, i.e., if there are N APs, each of them would have her specific $\lambda_i, i = 1,2, \dots N$ as well as common initial market conditions proxied by σ_X^2.

The second cost is the price impact cost incurred when the AP buys the underlying asset for creation. We assume that the price impact function is linear in the amount of the underlying asset bought for creation, $m(1-\gamma)X$, where $m > 0$ is known and exogenous.

Thus, the transaction price the AP faces when buying the underlying asset is $P_0 + m(1 - \gamma)X$. Intuitively, m reflects both informational and liquidity conditions in the underlying market irrespective of the individual AP, i.e., if there are M ETFs, each of them would have a specific price impact cost $m_j, j = 1,2, \dots M$ faced by any AP.

The AP's total payoff is given by

$$(P_e - P_0 - m(1-\gamma)X)(1-\gamma)X - \frac{\lambda\sigma_X^2}{2}(\gamma X)^2, \tag{1}$$

Under the assumption of perfect competition, the AP sets P_e to obtain zero profits

$$(P_e - P_0 - m(1-\gamma)X)(1-\gamma)X - \frac{\lambda\sigma_X^2}{2}(\gamma X)^2 = 0 \tag{2}$$

which, after re-arranging the terms is given by

$$P_e = \frac{\lambda\sigma_X^2 X}{2}\frac{\gamma^2}{1-\gamma} + P_0 + m(1-\gamma)X. \tag{3}$$

Furthermore, under perfect competition the representative AP chooses γ to offer the best price P_e to ETF investors. For $0 \leq \gamma < 1$, the FOC gives the following optimal choice of γ

$$\gamma^* = 1 - \sqrt{\frac{\lambda\sigma_X^2}{\lambda\sigma_X^2 + 2m}} \tag{4}$$

and the Second Order Condition is satisfied as $\frac{\lambda\sigma_X^2 X}{(1-\gamma)^3} > 0$.
For this γ^*, the equilibrium NAV^*, P_e^* and ETF mispricing are

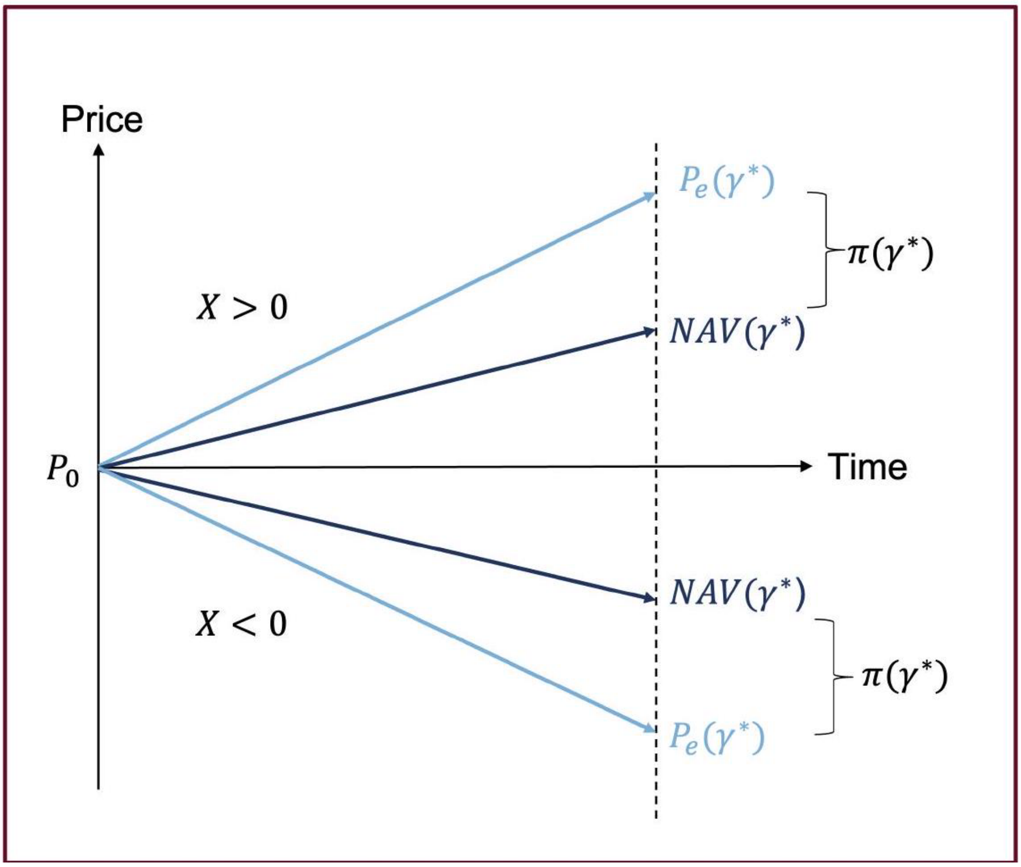
$$NAV^* = P_0 + mX\sqrt{\frac{\lambda\sigma_X^2}{\lambda\sigma_X^2 + 2m}}, \tag{5}$$

$$P_e^* = P_0 + \left(\sqrt{\lambda\sigma_X^2(\lambda\sigma_X^2 + 2m)} - \lambda\sigma_X^2\right)X, \tag{6}$$

$$\pi_e^* = P_e^* - NAV^* = \frac{(\gamma^*)^2}{1-\gamma^*}\frac{\lambda\sigma_X^2}{2}X. \tag{7}$$

Intuitively, when an ETF investor places a buy order in the market, the AP sells to the investor at a price P_e^* set to be higher than the NAV^*, translating into a positive mispricing. Conversely, when an ETF investor places a sell order in the market, the AP buys from the investor at a price P_e^* set to be lower than the NAV^*, translating into a negative mispricing. This key message is illustrated in Figure 3.

Figure 3: Illustration of ETF price and NAV dynamics



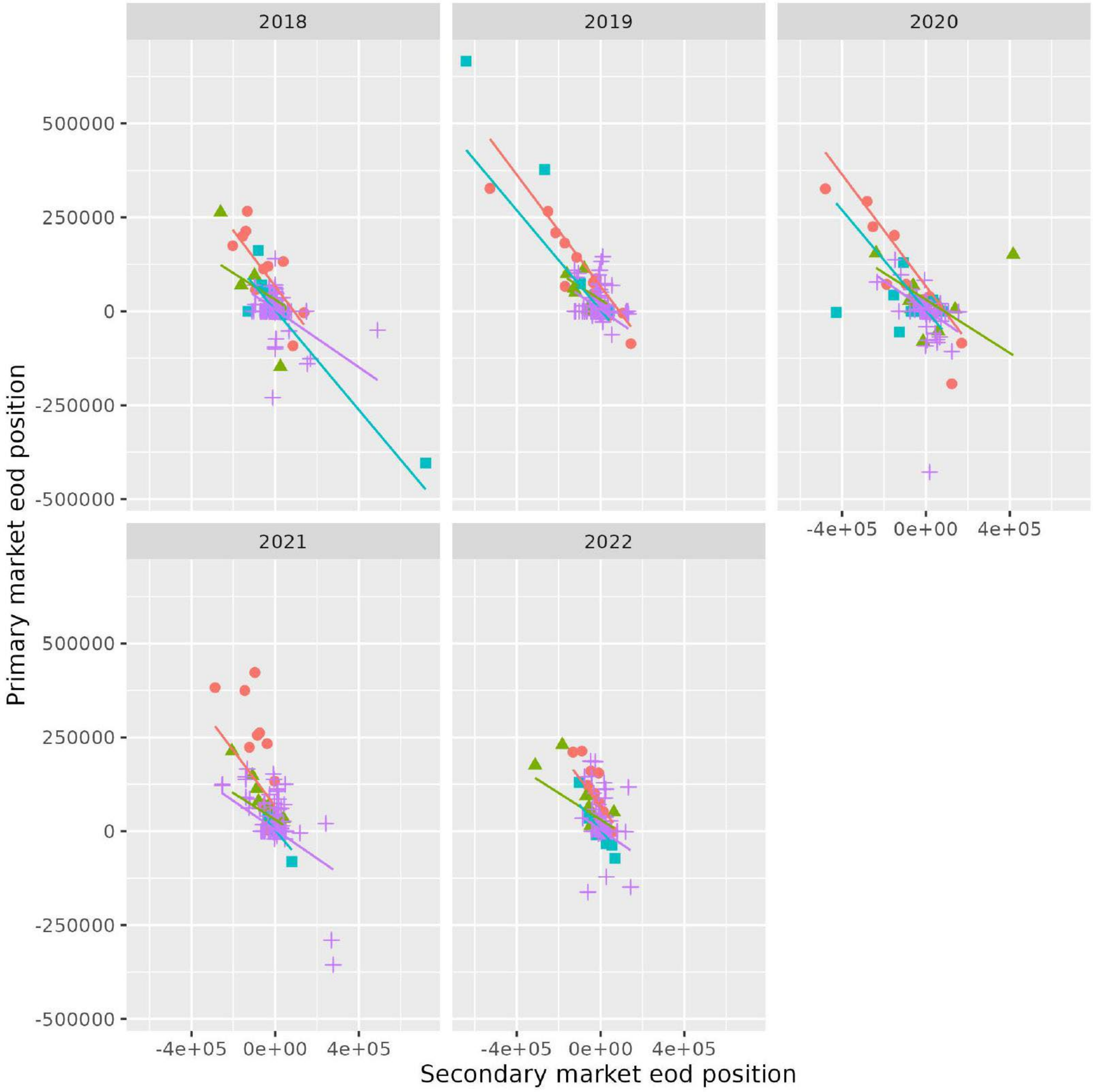
We can also notice that γ^* increases in m and decreases in λ. For a given $\lambda > 0$, if the price impact cost $m \rightarrow 0, \gamma^* \rightarrow 0$. Intuitively, when holding ETF inventory is costly to the AP, if the price impact cost in the market for the underlying is negligible, the AP prefers to hold a minimal amount of ETF inventory and use the creation/redemption functionality of the primary market for inventory management. For a given $m > 0, \lambda \rightarrow 0, \gamma^* \rightarrow 1$. Intuitively, when managing inventory in the market for the underlying is costly to the AP, if the cost of holding ETF inventory is negligible, the AP prefers to hold a maximal amount of ETF inventory and conduct inventory management in the secondary market.

From the static model, we learn from Equation 7 that $abs(\pi)$ increases with $\gamma, abs(X)$, and σ_X^2.

3.2 A dynamic model of AP's ETF trades

There are a few limitations of the static model that prevent us from deriving more realistic dynamics. Firstly, the model is static and only involves the AP's inventory decision. In reality, the AP trades multiple rounds with the market and therefore their behaviour in earlier rounds have implications for later rounds. As market makers, they not only make inventory but also pricing decisions. Secondly, the static model assumes that there's only one noise ETF investor in the market and thus the AP doesn't learn from observed order flow. Thirdly, the previous model assumes that there's only costs and no gains from taking on inventory, which diverges from the reality that the AP gains when the value of the held inventory increases. Also importantly, the static model restricts the inventory management parameter γ to be bounded between 0 and 1, which is not the case as can be observed from Figure 4. We plotted the representative APs' primary market EoD position against their secondary market EoD position and note that $\gamma = 0$ corresponds to a slope of -1 while $\gamma = 1$ correspond to a horizontal fitted line. It can be seen that there is heterogeneity between APs in terms of inventory management parameter γ and that γ doesn't always lie between 0 and 1.

Figure 4: APs' inventory management for iShares Core MSCI World UCITS ETF



Each point represents an AP's monthly-averaged positions in primary/secondary market. Blue squares, red circles and green triangles correspond to inventory positions of each of the top 3 APs and purple crosses represent all other APs. The slope of the fitted line for each AP describes their inventory management; a slope of -1 implies $\gamma = 0$.

To address the above-mentioned limitations, we extend the simple static model to a dynamic one. Our model builds on well-known models on market makers (see Madhavan and Smidt (1993)), with the addition that the AP has an alternative channel for inventory management: the ETF primary market. This also serves as an additional source of arbitrage opportunities.

The ETF market trades over multiple periods (days) $t = 1,2,\dots,T$, with one representative AP, one informed trader, and one noise trader. Let v_t represent the fundamental value of the ETF in period t. This fundamental value follows a random walk until period T, described by the equation $v_t = v_{t-1} + \eta_t$, where η_t is normally distributed with mean 0 and variance σ_η^2. On the final trading day T, the ETF pays a liquidating value equal to v_T. The value v_t is private information, observable only to informed traders at the beginning of each period. For uninformed agents, v_t is a random variable.

The final liquidation date is uncertain. Each day, there is a probability of $1-\rho$ that the day will be the liquidation date, i.e., $\Pr[t = T] = 1-\rho$.

Each day, traders submit their demand schedules to the AP. If that day is the liquidation date, all trades are cancelled, and the holdings of each trader are evaluated at their marked-to-market value. Otherwise, with probability ρ, trading occurs at a price P_t, set by the risk-neutral AP after observing the aggregate excess demand schedule. Let $Z_t(P_t)$ denote the excess demand schedule at price P_t. This excess demand originates from both informed and noise traders. Therefore, we can decompose Z_t as $Z_t = Q_t + X_t$, where $Q_t = \delta(v_t - P_t)$, with $\delta > 0$ representing the quantity demanded by informed traders, and X_t being the quantity demanded by noise traders. The former increases with the difference between the fundamental value and price, while the latter is a normal random variable with mean 0 and variance σ_X^2.

The AP observes the aggregate excess demand schedule

$$Z_t(P_t) = Q(v_t, P_t) + X_t = \delta(v_t - P_t) + X_t = D_t - \delta P_t \tag{8}$$

before setting the price, where $D_t = \delta v_t + X_t$ is the intercept of the demand schedule. Let $\mu_t = E[v_t \mid \Phi_t]$ be the conditional expectation of the fundamental value given the information set Φ_t of the AP at period t (which includes both current and past trading history), after observing the aggregate excess demand schedule. The AP forms this expectation by observing a noisy signal

$$d_t = \frac{D_t}{\delta} = v_t + \frac{X_t}{\delta} \tag{9}$$

around it (observing the excess demand schedule is equivalent to observing its intercept and its slope) with an error variance of $\sigma_d^2 = \frac{\sigma_X^2}{\delta^2}$.

Along with the knowledge that v_t follows a random walk, the AP can form their expectation of the fundamental value μ_t by using the steady-state Kalman Filter updating equation[2]

[2] The Kalman filter updating equation is derived from the dynamics:

$$\mu_t = \Omega d_t + (1-\Omega)\mu_{t-1}, \tag{10}$$

where Ω depends on the signal-to-noise ratio $\Upsilon = \frac{\sigma_\eta^2}{\sigma_d^2}$ such that

$$\Omega = \frac{-\Upsilon + \sqrt{\Upsilon^2 + 4\Upsilon}}{2}. \tag{11}$$

With this understanding of the fundamental value, the AP can form their belief about the quantity of informed trader demand, $\delta(\mu_t - P_t)$, and noise trader demand, x_t. Both of these are unbiased estimates of the true informed trader and noise trader's demand.

The AP's inventory fluctuates as a result of their market-making and arbitrage activities. Let I_t denote the AP's inventory at the beginning of trading day t, and let γ represent the proportion of market demand that the AP satisfies using their existing inventory (instead of going to the ETF primary market). The inventory evolves according to the following equation

$$I_t = I_{t-1} - \gamma_{t-1} Z_{t-1}, \tag{12}$$

where Z_t is the aggregate excess demand at time t.
The AP's wealth at the beginning of the period is the sum of the value of their inventory, the opening capital, and the value of outside assets that generate external income. This is expressed as

$$W_t = v_t I_t + K_t + y_t, \tag{13}$$

where K_t represents the capital at the beginning of the period and y_t is the outside income from other investments with variance σ_y^2.

At the start of each period, the AP incurs an inventory cost that depends on the variance of their beginning wealth, given by

$$c_t = \omega\sigma^2(W_t), \tag{14}$$

where $\omega > 0$ is the inventory cost per unit of variance. This cost can be interpreted as the funding cost faced by the AP, which scales with the riskiness of their current wealth.

By substituting the formula for W_t, the inventory $\text{cost} c_t$ can be further written as

$$c_t = \omega\big(\phi_0 + \phi_1(I_t - I^d)\big), \tag{15}$$

where the optimal inventory level is $I^d = -\frac{\sigma_{vy}}{\sigma_v^2}, \phi_0 = \sigma_y^2 - \left(\frac{\sigma_{vy}}{\sigma_v^2}\right)^2$, and $\phi_1 = \sigma_v^2$ where σ_v is the conditional variance of v_t before the information at period t and σ_{vy} is the covariance

$$d_t = v_t + \frac{X_t}{\delta}, v_t = v_{t-1} + \eta_t$$

The steady state refers to as time goes to infinity (after many rounds of observing the signal and updating their beliefs), the variance around the belief converges to a constant. The coefficient Ω (also known as the Kalman gain) is no longer time-varying and becomes a function of the signal-to-noise ratio Υ. See Chapter 2 in Durbin and Koopman (2012) for more details.

between v_t and y_t. Notice that the AP's optimal ETF inventory I^d is positive when the covariance with outside income is negative as the ETF can be used for hedging purposes.

The AP's capital evolves according to the following rule

$$K_t = K_{t-1} - c_{t-1} - y_{t-1} + P_{t-1}Z_{t-1} - (1-\gamma)Z_{t-1}P_t^u, \tag{16}$$

where $K_{t-1} - c_{t-1} - y_{t-1}$ are capital carried over from yesterday. $P_{t-1}z_{t-1}$ is the amount the AP receives from selling ETFs and $(1-\gamma)z_{t-1}P_t^u$ is the amount the AP pays to buy the underlying assets for ETF creation.

We assume that the NAV only changes when the AP trades in the primary market, and as in the static model, we further assume that the AP does not carry any inventory in the underlying assets. Therefore, if the AP needs to create ETF shares, they must buy all the underlying assets from the underlying market, and conversely, sell all underlying assets when redeeming shares.

As a result, the NAV follows the following dynamic

$$P_t^u = P_{t-1}^u + m(1-\gamma_{t-1})Z_{t-1}, \tag{17}$$

where $m > 0$ represents the price impact cost the AP faces when trading in the primary market.

The AP's objective is to maximize final period wealth, with the timing of the final period being uncertain. The AP achieves this by choosing ETF prices P_t and the parameter γ_t each period as a function of the observed information

$$\max_{P_t,\gamma_t} \sum_{j=t}^{\infty} E\left[W_j\right] \cdot \Pr[j = T], \tag{18}$$

where W_j represents the wealth at period j and $\Pr[j = T]$ is the probability that period j is the final period.

To simplify the problem, we use the Bellman equation, transforming this multi-period problem into a recursive sequence of two-period problems. Each period, the AP chooses their decision variables to maximize both the current and expected future payoffs, conditioned on the available state variables. The decision made in the current period affects the next period's state variables. In this way, the optimal choice rule becomes a function of the state variables, and we can derive the optimal dynamics for these variables.

At the beginning of each period, the information available to the AP includes the opening ETF inventory, opening capital, the last period's NAV, and the excess demand schedule. Instead of using the excess demand schedule Z_t directly as a state variable, it is more convenient to decompose it based on the AP's belief about demand from informed traders and noise traders. The predicted demand from informed traders is $\delta(\mu_t - P_t)$, and the predicted demand from noise traders is x_t.

Therefore, the state variables for decision-making in period t are:

$$I_t, K_t, x_t, \mu_t, P_t^u$$

where I_t is the ETF inventory, K_t is the capital, x_t is the noise trader demand, μ_t is the AP's belief about the fundamental value, and P_t^u is the NAV of the underlying assets.

We can therefore express the AP's problem as the solution of the following Bellman equation (where the ' denotes the next period's state variables, to simplify notation)

$$J(I, K, x, \mu, P^u) = \max_{p,\gamma} E\left[(1-\rho)(\mu I + K - c) + \rho J\left(I', K', x', \mu', P^{u'}\right) \mid \Phi\right]. \tag{19}$$

Based on the updating rules for K, I, and P^u explained in Equations 12, 16, and 17, we can formulate the AP's expectation of the following period's state variables based on current information

$$E[I' \mid \Phi] = I - \gamma z = I - \gamma\{\delta(\mu - P) + x\}, \tag{20}$$

$$\begin{aligned} E[K' \mid \Phi] &= K - c + y + Pz - (1-\gamma)z\{P^u + m(1-\gamma)z\} \\ &= K - c + y + \{P - (1-\gamma)P^u\}z - m(1-\gamma)^2 z^2, \end{aligned} \tag{21}$$

$$E[x' \mid \Phi] = 0, \tag{22}$$

$$E[\mu' \mid \Phi] = \mu, \tag{23}$$

$$E\left[P^{u'} \mid \Phi\right] = P^u + m(1-\gamma)z. \tag{24}$$

The First Order Conditions of the current value function $J(I, K, x, \mu, P^u)$ with respect to the two choice variables P and γ are

$$\begin{aligned} \frac{\partial J}{\partial P} = \quad & \delta\gamma E[U_l'] + \{\delta\mu - 2\delta P + x + (1-\gamma)P^u\delta + 2m\delta(1-\gamma)^2 z\}E[U_K'] \\ & -m(1-\gamma)\delta E\left[U_{p^u}'\right] = 0, \end{aligned} \tag{25}$$

$$\frac{\partial J}{\partial \gamma} = -zE[J_I'] + \{P^u z + 2m(1-\gamma)z^2\}E[J_K'] - mzE[J_{P^u}'] = 0. \tag{26}$$

The Envelope conditions are derived by differentiating the value function $J(I, K, x, \mu, P^u)$ with respect to each of the state variables. This gives

$$J_I = (1-\rho)\mu - \left[(1-\rho) + \rho E[J_K']\right](2\omega\phi_1)(I - I^d) + \rho E[J_I'], \tag{27}$$

$$J_K = (1-\rho) + \rho E[J_K'], \tag{28}$$

$$J_x = \rho\{-\gamma E[J_I'] + [P - (1-\gamma)P^u - 2m(1-\gamma)^2 z]E[J_K'] + m(1-\gamma)E[U_{P^u}']\}, \tag{29}$$

$$J_\mu = (1-\rho)I + \rho\{-\gamma\delta E[J_I'] + [(P - (1-\gamma)P^u)\delta - 2m\delta(1-\gamma)^2 z]E[J_K'] + E\left[J_\mu'\right] + m(1-\gamma)\delta E[J_{P^u}']\}, \tag{30}$$

$$J_{P^u} = \rho\{-(1-\gamma)zE[J_K'] + E[J_{P^u}']\}. \tag{31}$$

Based on the FOCs and Envelope conditions, we make the following guess for the functional form of the value function

$$\begin{aligned} J(I,K,x,\mu,P^u) = \text{const } &+A_0\mu I + K + A_1(I-I^d)^2 + A_2(I-I^d)x + A_3x^2 \\ &+A_4P^u\mu + A_5P^u(I-I^d) + A_6P^ux + A_7(P^u)^2. \end{aligned} \tag{32}$$

The guessed form of the value function yields the following expectations of its derivative with respect to each of the state variables

$$E[J'_I] = A_0\mu + 2A_1(I-I^d-\gamma z) + A_5(P^u + m(1-\gamma)z), \tag{33}$$

$$E[J'_{P^u}] = A_4\mu + A_5(I-I^d-\gamma z) + 2A_7(P^u + m(1-\gamma)z), \tag{34}$$

$$E[J'_x] = A_2(I-I^d-\gamma z) + A_6(P^u + m(1-\gamma)z), \tag{35}$$

$$E\left[J'_\mu\right] = A_0(I-\gamma z) + A_4(P^u + m(1-\gamma)z), \tag{36}$$

$$E[J'_{P^u}] = A_4\mu + A_5(I-I^d-\gamma z) + 2A_7(P^u + m(1-\gamma)z). \tag{37}$$

We can then substitute these expressions back into the FOCs and envelope conditions to write J_I, J_K, J_x, J_μ, and J_{P^u} as functions of the state variables. By equating the coefficients from these derived functions with the coefficients from the direct derivative of the hypothesized value function (Equation 32) with respect to the state variables, we obtain the following solutions for our parameters $A_0, A_1, A_2, A_3, A_4, A_5, A_6$, and A_7.

Initially, we find four sets of solutions that satisfy all the conditions. However, for an economically meaningful solution (i.e., a deviation from the optimal level of inventory should decrease the value function), we impose the restriction $A_1 < 0$ for any $d > 0, m > 0$, $0 < r < 1, \phi > 0$, and $\omega > 0$. This restriction rules out three of the solutions, leaving us with the following solution for A_1

$$\begin{aligned} A_1 = &\frac{-2A\delta G + (-1+\rho)^2(-1+G+\rho)(-m(-1+\rho)^2 - 2\phi\omega(1-\rho+\delta m(1+\rho)))}{2\delta(-1+\rho)^2(m(-1+\rho)^2 + \phi(2-2\rho+\delta m(3+\rho))\omega)} \\ &-(-1+\rho)^2\frac{-\delta\phi^2\omega^2\big(2-2\rho+\delta m(3+G+\rho)\big)}{2\delta(-1+\rho)^2(m(-1+\rho)^2 + \phi(2-2\rho+\delta m(3+\rho))\omega)}, \end{aligned} \tag{38}$$

where
$A = \sqrt{\phi^2(-1+\rho)^3\omega^2(-m^2(-1+\rho)^2 - 2m\phi(1-\rho+\delta m(1+\rho))\omega + \phi^2(-(1+\delta m)^2+\rho)\omega^2)}$,
and

$$\begin{aligned} G^2 = &\frac{m^2(-1+\rho)^4 + 2m\phi(-1+\rho)^2(2-2\rho+\delta m(1+\rho))\omega}{4\phi^2\omega^2 + 4m\phi\omega(1-\rho+\delta\phi\omega) + m^2(-1+\rho+\delta\phi\omega)^2} \\ &+\frac{\phi^2\left(\delta m\big(-8+\delta m(-5+\rho)\big) + 4(-1+\rho)\right)(-1+\rho)\omega^2}{4\phi^2\omega^2 + 4m\phi\omega(1-\rho+\delta\phi\omega) + m^2(-1+\rho+\delta\phi\omega)^2} \\ &+\frac{4\delta mA}{4\phi^2\omega^2 + 4m\phi\omega(1-\rho+\delta\phi\omega) + m^2(-1+\rho+\delta\phi\omega)^2}, \end{aligned} \tag{39}$$

and $G > 0$.
It can also be shown that[3] $1 - \rho < G < 1 - \rho + dm$ for any $d > 0, m > 0, 0 < \rho < 1$, $\phi > 0$, and $\omega > 0$, which results in $A_1 < 0$.

The rest of the parameters take the following form

$$A_0 = 1 - \frac{-1 + G + \rho}{dm} \in (0,1), \tag{40}$$

$$A_2 = -A_1 - \phi\omega \in (-\phi\omega, \infty), \tag{41}$$

$$A_3 = \frac{1}{4}\left(\frac{\rho}{\delta} + A_1 + \phi\omega\right) \in (0, \infty), \tag{42}$$

$$A_5 = \frac{-1 + G + \rho}{dm} \in (0,1), \tag{43}$$

$$A_6 = -\frac{-1 + G + \rho}{2dm} \in \left(-\frac{1}{2}, 0\right), \tag{44}$$

$$A_7 = -\frac{(\rho - 1)A_1}{2\omega\phi m} + \frac{(-1 + G + \rho)\big(m(1 - \rho) - \phi\omega(2 + dm)\big)}{2dm(2\omega\phi m)} + \frac{1}{2m} \in \left(0, \frac{1}{m}\right), \tag{45}$$

$$A_4 = -2A_7 \in \left(-\frac{2}{m}, 0\right). \tag{46}$$

The equilibrium price as a function of the state variables is as follows

$$P^* = \mu + \alpha_1(I - I^d) + \alpha_2 x + \alpha_3(P^u - \mu), \tag{47}$$

where

$$\alpha_1 = -\frac{m\big(A_5^2 m + A_1(4 - 4A_7 m)\big)}{A_1\big(4 - 4\delta m(-1 + A_7 m)\big) + m\big(-4 + 4A_7 m + A_5(4 + A_5\delta m)\big)} < 0, \tag{48}$$

$$\alpha_2 = \frac{A_1\big(-2 + 4\delta m(-1 + A_7 m)\big) - m\big(-2 + 2A_7 m + A_5(2 + A_5\delta m)\big)}{\delta\Big(4A_1\big(-1 + \delta m(-1 + A_7 m)\big) - m\big(-4 + 4A_7 m + A_5(4 + A_5\delta m)\big)\Big)} > 0, \tag{49}$$

$$\alpha_3 = \frac{(-1 + A_5)A_5 m + A_1(2 - 4A_7 m)}{A_1\big(4 - 4\delta m(-1 + A_7 m)\big) + m\big(-4 + 4A_7 m + A_5(4 + A_5\delta m)\big)} > 0. \tag{50}$$

The parameter α_1 is negative and increases with demand elasticity δ, decreases with the inventory cost parameter $\phi\omega$, and decreases with the price impact m. This shows that the AP offers more attractive prices for selling ETFs when their inventory exceeds the optimal level. This effect is stronger when ETF investors have less elastic demand, and when holding inventory becomes more expensive or managing inventory via redemption becomes more costly. The parameter α_2 is positive, meaning that the noise trader demand pushes up the equilibrium price. This effect decreases with demand elasticity δ, and increases with inventory cost $\phi\omega$ and the price impact parameter m. The AP faces a trade-off between charging higher prices to recover the cost of managing inventory while

[3] For all the sign checks in the model results, we used Mathematica to simulate various combinations of base parameter values, and the stated signs hold in all experimented cases, so we believe they are generically true.

avoiding a decrease in the informed trader's demand. The parameter α_3 is positive, implying that the larger the deviation between NAV and the fundamental value, the further the ETF price deviates from the fundamental value. This shows that the AP's arbitrage activities between the ETF and the underlying market help keep the two prices aligned, even if both deviate significantly from the fundamental value. This co-movement effect decreases with the demand elasticity δ and the price impact parameter m, but increases with the inventory cost parameter $\phi\omega$. This is intuitive, as the prior two effects lead the AP to base their pricing decision more on the ETF secondary market (either when the secondary market demand is more sensitive to price change or because the primary market is less accessible), whereas the increase in inventory cost incentivizes inventory management through the primary market, strengthening the co-movement between the NAV and the ETF price.

The ETF price is centred around a weighted average (note that $\alpha_3 \in (0,1)$) of fundamental value and NAV and only deviates from this level when either inventory deviates from optimal or predicted noise trader demand deviates from 0. This is an extension to the static model. From Equation 3, we can see that the ETF price is centred around the starting fundamental value and deviation is caused by exogenous shocks to market demand. Extending the model to a dynamic framework allows the ETF price to reflect both the previous period NAV (for arbitrage considerations) and the evolving fundamental value (which results from the market maker's learning of excess demand schedule over time). When α_3 goes to 0 and the inventory is at the optimal level, the ETF pricing decision reverts back to that implied by the static model.

The optimal γ as a function of the state variables is as follows

$$\gamma^* = \frac{\beta_1(I - I^d) + \beta_2 x + \beta_3(P^u - \mu)}{\zeta_1(I - I^d) + \zeta_2 x + \zeta_3(P^u - \mu)}, \tag{51}$$

where

$$\beta_1 = A_5 m(2 + A_5\delta m) + A_1\big(4 - 4\delta m(-1 + A_7 m)\big) < 0, \tag{52}$$

$$\beta_2 = m(-2 + A_5 + 2A_7 m) < 0, \tag{53}$$

$$\beta_3 = -2 + 4A_7 m + A_5(2 + \delta m) < 0, \tag{54}$$

$$\zeta_1 = \delta m\big(A_5^2 m + A_1(4 - 4A_7 m)\big) < 0, \tag{55}$$

$$\zeta_2 = 2\big(A_1 + m(-1 + A_5 + A_7 m)\big) < 0, \tag{56}$$

$$\zeta_3 = \delta\left(-\big((-1 + A_5)A_5 m\big) + A_1(-2 + 4A_7 m)\right) > 0. \tag{57}$$

Given the optimal ETF prices, the optimal proportion (γ) of ETF demand that APs satisfy with inventory ensures that the marginal effect of keeping ETFs in inventory is balanced with the marginal effect of creating/redeeming them. The optimal γ in Equation 51 shows that it takes the form of a ratio of two linear combinations of the state variables: $I - I^d, x$, and $P^u - \mu$.

To facilitate understanding, we can interpret γ^* when each of the state variables approaches its limit or dominates using L'Hôpital's rule and the implications highlight 3 different aspects of the AP's behaviour: inventory management, market making and arbitrage.

As $I - I^d \to \infty$ (Inventory management):

$$\gamma = \frac{\beta_1}{\zeta_1} > 1 \tag{58}$$

From the optimal price rule 47, we can see that the optimal theoretical price tends to $-\infty$, resulting in market excess demand $z \to \infty$ from Equation 8. In this scenario, the AP would sell ETFs to satisfy investors while redeeming existing ETF inventory to reduce it back to the optimal level. The term $\gamma = \frac{\beta_1}{\zeta_1}$ increases with the inventory cost parameter $\phi\omega$ and decreases with demand elasticity δ and the price impact parameter m. The higher the inventory cost, the greater the AP's incentive to reduce current inventory by redeeming ETFs in the primary market, in addition to selling in the secondary market. However, if demand is elastic or the cost of redemption is high, the AP would prefer to clear their inventory in the ETF secondary market.

As $x \to \infty$ (Market making):

$$0 < \gamma = \frac{\beta_2}{\zeta_2} < 1 \tag{59}$$

The optimal theoretical price tends to ∞, resulting in excess demand from informed traders going to $-\infty$. This leads to a reasonable level of total market excess demand (because $x \to \infty$). The AP would buy or sell ETFs, satisfying demand with its inventory and the ETF primary market. The term $\gamma = \frac{\beta_2}{\zeta_2}$ decreases with the inventory cost parameter $\phi\omega$, while it increases with the price impact parameter m and demand elasticity δ. This is intuitive: a higher inventory cost leads to more demand being satisfied via the primary market, while an increase in price impact or a more elastic secondary market means it is cheaper to manage inventory using the ETF secondary market.

As $P^u - u \to \infty$ (Arbitrage):

$$\gamma = \frac{\beta_3}{\zeta_3} < 0. \tag{60}$$

The optimal theoretical price tends to ∞, and the market excess demand tends to $-\infty$. In this case, the AP buys ETFs from the secondary market while redeeming even more than the amount they purchase. This is due to the AP taking advantage of the high NAV price to obtain arbitrage profits. The term $\gamma = \frac{\beta_3}{\zeta_3}$ increases (proportion of extra redemption decreases) with demand elasticity δ and the inventory cost parameter $\phi\omega$, while its behaviour with respect to the price impact parameter m is mixed.

As the inventory cost increases, it becomes less preferable to move existing inventory for arbitrage purposes. With more elastic ETF demand, the AP can buy a large number of

ETFs without significantly raising the purchasing price, making it unnecessary to adjust their existing inventory for arbitrage. However, as the price impact parameter increases, the AP offers lower purchasing prices in the ETF market according to Equation 50, thus reducing the amount of ETFs bought from the secondary market. At the same time, as the price impact parameter increases, it becomes less profitable to arbitrage using the AP's own inventory, leading to mixed results.

The above scenarios demonstrate that the γ in the dynamic model no longer is restricted between 0 and 1 but responds to both AP's inventory management and arbitrage incentives.

With the optimal choice variables P^* and γ^*, we derive the optimal dynamics for the state variables, which are the testable implications from our model, and we estimate them using real data in the Section 6.

Inventory dynamic

$$I' = I - \iota_1(I - I^d) - \iota_2 x - \iota_3(P^u - \mu), \tag{61}$$

where

$$\iota_1 = \frac{A_5 m(2 + A_5\delta m) + A_1\big(4 - 4\delta m(-1 + A_7 m)\big)}{A_1\big(4 - 4\delta m(-1 + A_7 m)\big) + m\big(-4 + 4A_7 m + A_5(4 + A_5\delta m)\big)} > 0, \tag{62}$$

$$\iota_2 = \frac{-2m + A_5 m + 2A_7 m^2}{A_1\big(4 - 4\delta m(-1 + A_7 m)\big) + m\big(-4 + 4A_7 m + A_5(4 + A_5\delta m)\big)} > 0, \tag{63}$$

$$\iota_3 = \frac{2A_5 + 2(-1 + 2A_7 m)}{A_1\big(4 - 4\delta m(-1 + A_7 m)\big) + m\big(-4 + 4A_7 m + A_5(4 + A_5\delta m)\big)} > 0. \tag{64}$$

The derived inventory is a mean-reverting process around the optimal level of inventory I^d, as $\iota_1 \in (0,1)$, keeping fixed the predicted noise trader demand and the NAV's deviation from the fundamental value. The speed of adjustment, represented by ι_1, increases with demand elasticity δ and the inventory cost parameter $\phi\omega$. It decreases with the price impact parameter m. As demand becomes more elastic, inventory reverts faster with a given level of price change. The AP is also more willing to keep the inventory level close to the optimal when inventory is more costly. However, if the option of going to the primary market to offload inventory becomes more expensive, inventory management becomes more difficult, and the inventory level reverts more slowly to the optimal level.

In addition, the AP's inventory decreases when market demand increases and responds negatively to the deviation of NAV from the fundamental value. These behaviours are driven by the AP's market-making and arbitrage functions.

Price dynamic

$$\begin{aligned} P^{*'} - P^* &= (\mu' - \mu) + \alpha_1(I' - I) + \alpha_2(x' - x) + \alpha_3\left(P^{u'} - \mu' - (P^u - \mu)\right) \\ &= (1 - \alpha_3)(\mu' - \mu) + \alpha_1(I' - I) + \alpha_2(x' - x) + \alpha_3\big(P^{u'} - P^u\big) \end{aligned} \tag{65}$$

Using the definitions for μ in Equation 10 and the definition for d_t, we can rewrite the changes in the fundamental value as[4]

$$\mu' - \mu = \frac{\Omega x'}{\delta(1-\Omega)}.$$

Substituting this into the above price updating equation gives

$$P' - P = (1-\alpha_3)\frac{\Omega}{\delta(1-\Omega)}x' + \alpha_1(I'-I) + \alpha_2(x'-x) + \alpha_3\left(P^{u'} - P^u\right). \tag{66}$$

The AP's price revision is a linear combination of changes in the predicted fundamental value, changes in the inventory level, changes in the predicted noise trader demand, and changes in the NAV. The changes in the predicted fundamental value arise from the update after observing the difference between the excess demand and the predicted informed trader's demand.

The first term on the RHS of Equation 66 represents an information effect on prices, and this effect becomes more pronounced as market demand becomes more informative (i.e., as Ω approaches 1). As the NAV that the AP gets charged increases, the AP will also charge a higher price for selling the ETF due to arbitrage. However, since $\alpha_3 \in (0,1)$, the price update only partially responds to changes in the NAV. Finally, when the AP's inventory level is higher, the AP will charge a lower price to offload unwanted inventory.

NAV dynamic

$$P^{u'} = P^u - 2\theta_1(I - I^d) + \theta_2(P^u - \mu) + \theta_1 x, \tag{67}$$

where

$$\theta_1 = m\frac{2A_1 + A_5 m}{A_1\left(4 - 4\delta m(-1 + A_7 m)\right) + m\left(-4 + 4A_7 m + A_5(4 + A_5\delta m)\right)} > 0, \tag{68}$$

$$\theta_2 = -1 + \frac{-4A_1 - 2(-1 + A_5 + A_1\delta)m}{4A_1\left(-1 + \delta m(-1 + A_7 m)\right) - m\left(-4 + 4A_7 m + A_5(4 + A_5\delta m)\right)} < 0. \tag{69}$$

The NAV updates according to the APs' inventory level. If the AP's current inventory becomes excessive, they are more likely to use the primary market to offload some inventory, which decreases the NAV. This effect is stronger when the demand elasticity δ is lower, and when either the inventory cost $\phi\omega$ or the price impact parameter m is higher. This is intuitive because when holding inventory brings significant disutility and managing it through changing prices becomes difficult, the AP will turn to the primary market. With a higher price impact parameter, the NAV will be reduced further.

The NAV also tends to self-correct, moving closer to the predicted fundamental value. This is due to the AP arbitraging between the two markets, demonstrating the ETF's role as an important tool for price discovery. The speed of correction $\theta_2 \in \left(-\frac{1}{2}, 0\right)$ indicates

[4] The proof for this is as follows. From Equation 10, we know that

$$\mu' - \mu = \Omega(d' - \mu) = \Omega\left(\mu' + \frac{x'}{\delta} - \mu\right) = \frac{x'}{\delta}\Omega + \Omega(\mu' - \mu)$$

The result follows by rearranging.

that the deviation from the predicted fundamental value could be persistent and only corrected over time. This speed is faster with a higher m, a lower $\phi\omega$, and a higher δ. A larger price impact parameter escalates the effect of trading in the primary market. Lower penalties for deviation from optimal inventory and more elastic demand both make it easier for the AP to acquire shares to conduct arbitrage, facilitating faster correction of the NAV.

Mispricing dynamic

$$\begin{aligned} \pi' &= \tau_1\pi + \tau_2(I' - I^d) + \tau_3(I - I^d) + \tau_4 x' + \tau_5 x + \tau_6(\mu' - \mu) \\ &= \tau_1\pi + \tau_2(I' - I^d) + \tau_3(I - I^d) + \tau_4 x' + \tau_5 x + \tau_6 \frac{\Omega}{\delta(1-\Omega)} x', \end{aligned} \tag{70}$$

where

$$\tau_1 = 1 + \theta_2 > 0, \tag{71}$$

$$\tau_2 = \alpha_1 + 2\theta_1, \tag{72}$$

$$\tau_3 = \alpha_3(-2\theta_1) - \alpha_1(1 + \theta_2) > 0, \tag{73}$$

$$\tau_4 = \alpha_2 - \theta_1 > 0, \tag{74}$$

$$\tau_5 = \alpha_3\theta_1 - \alpha_2(1 + \theta_2) < 0, \tag{75}$$

$$\tau_6 = 1 - \alpha_3 + \theta_2 > 0. \tag{76}$$

We define mispricing as the difference between the current ETF price and the next period's NAV, $\pi = P - P^{u'}$. This is because the NAV P^u in our model is the beginning-of-period NAV and only incorporates information from the previous period. Anecdotal evidence suggests that APs usually immediately lock in arbitrage profits once they detect a mispricing by longing/shorting the ETF and shorting/longing the underlying asset and use creation and redemption at the end of the day to unwind their positions. Therefore, both the closing ETF price and the NAV should incorporate demand information from the day, corresponding to P and $P^{u'}$ in our model.

The effect of the current period inventory on mispricing is mixed. A higher inventory level reduces both the ETF price and the next period's NAV, and the aggregate effect depends on the relative magnitudes of α_1 and $-2\theta_1$, as seen in the equations for optimal price (Equation 47) and optimal NAV dynamics (Equation 67). When the inventory cost $\phi\omega$ is higher, the AP opts to use the primary market for inventory management, reducing the NAV more, resulting in a more positive effect on mispricing. Conversely, when the price impact cost m is higher, the AP reduces the price more to decrease inventory, resulting in a more negative effect. As demand becomes more elastic, the effect on mispricing first increases (potentially from negative to positive) and then decreases (towards zero). The initial increase occurs because the price drops less when selling ETF inventory, while the subsequent decrease results from a shift from selling in the underlying market to selling in the ETF market, reducing the drop in NAV.

The effect of lagged inventory and noise trader demand on current mispricing arises from their impact on NAV. Since the NAV accumulates and changes according to the price impact in each period, a higher lagged inventory and lower lagged noise trader demand depress future periods' NAV, thereby increasing mispricing.

The effects of current noise trader demand and updates in the fundamental value on mispricing are both positive. Although they have positive effects on both the price and the NAV, the effect on the price is stronger because prices respond instantaneously, while the NAV only adjusts when the AP resorts to the primary market for inventory management. This resonates with the message in the static model and Figure 3.

Importantly, from Equation 70 and knowing that $\tau_1 \in \left(\frac{1}{2}, 1\right)$, we can see that mispricing mean-reverts toward 0 at a slow rate (keeping other variables fixed) but fluctuates with predicted noise trader demand when the AP's inventory is kept at the optimal level. The mean reversion is faster (τ is smaller) when demand is more elastic, inventory costs are lower, and price impact costs are higher.

To facilitate understanding, imagine the ETF price is below the NAV, and the AP wants to buy the ETF, redeem it, and sell the underlying. More elastic demand implies that it doesn't take a large price movement for the AP to purchase ETF shares from the secondary market. This encourages the AP's arbitrage activity, as it helps maintain a larger profit margin. A higher price impact cost implies that selling in the underlying market would depress the NAV more, speeding up the mispricing correction. With lower inventory costs, the AP would willingly deviate from their optimal inventory level to profit from arbitrage.

The informed trader's problem

For simplification, we justify the demand function of the informed trader by assuming they are myopic and unable to access any information about the ETF primary market, including the AP's actions and the NAV. Therefore, they trade only in the ETF secondary market and consider the AP as a conventional market maker. The model then reverts back to that in Madhavan and Smidt (1993), corresponding to the case where $\gamma = 1$, and $A_4 = A_5 = A_6 = A_7 = 0$. Using the argument in Madhavan and Smidt (1993), one can show that the AP's assumed informed trader's demand function $\delta(v_t - P_t)$ is optimal for such a myopic informed trader. This is achieved by demonstrating that any deviation from $\delta(v_t - P_t)$ is suboptimal.

Time-varying optimal level of inventory

We can introduce a time-varying optimal level of inventory similarly to Madhavan and Smidt (1993). Since the optimal level of inventory is $I^d = -\frac{\sigma_{vy}}{\sigma_v^2}$, it is reasonable to assume that the time-varying I^d is driven mostly by changes in the covariance of the fundamental value with outside income, σ_{vy}. We assume that this covariance follows a random walk, $\sigma_{vy}' = \sigma_{vy} + \pi$, where π is a mean 0 random variable. We now introduce a new state variable I^d into our previous model, with $E\left[I^{d'} \mid \Phi\right] = I^d$. Our new Bellman equation becomes

$$J^*(I, K, x, \mu, P^u, I^d) = \max_{p,\gamma} E\left[(1-\rho)(\mu I + K - c) + \rho J\left(I', K', x', \mu', P^{u'}, I^{d'}\right) \mid \Phi\right]. \tag{77}$$

Our new guessed form of the value function is

$$\begin{aligned} J^*(I, K, x, \mu, P^u, I^d) = \text{const} + A_0 \mu I + K + A_1 (I - I^d)^2 + A_2 (I - I^d)x + A_3 x^2 \\ + A_4 P^u \mu + A_5 P^u (I - I^d) + A_6 P^u x + A_7 (P^u)^2. \end{aligned} \tag{78}$$

Since the optimal level of inventory follows a random walk, it does not affect the optimal choice variables as it is not possible to hedge the associated risk. However, the time-varying optimal inventory reduces the expected value function to account for the additional risk, as follows

$$\begin{aligned} E\left[\left(I' - I^{d'}\right)^2 \mid \Phi\right] &= \text{Var}\left[I' - I^{d'} \mid \Phi\right] + E\left[I' - I^{d'} \mid \Phi\right]^2 \\ &= \text{Var}\left[\frac{-\pi}{\sigma_v^2}\right] + (I' - I^d)^2. \end{aligned} \tag{79}$$

Therefore,

$$E\left[J^{*'} \mid \Phi\right] = E[J' \mid \Phi] + A_1 \text{Var}\left[\frac{-\pi}{\sigma_v^2}\right] < E[J' \mid \Phi].$$

All the dynamics in the baseline model remain the same once we use the realized time-varying I^d to replace the original static I^d. The equilibrium demand function of the informed traders would also remain unaffected, as changes in I^d are unpredictable and thus cannot be factored into their expectations of future prices.

4 Data

4.1 Primary market data

We obtained our primary market data from two major ETF issuers, which include daily creation and redemption activities of all active APs for each ETF in our sample from January 2018 to October 2022. The fields relevant to our study are the trade date, the AP LEI number and name, the price and quantity of ETF shares transacted, the trade type (whether they are in-kind or in-cash), and the types of APs they identify themselves as (either Investment/Wholesale Bank or Broker-Dealer/Market Maker).

Compared to the existing N-CEN reporting dataset, which records the annual total activities of each AP, our analysis is based on more granular daily data. It is also worth noting that our dataset distinguishes between primary market trades made by the APs themselves and those made on behalf of their clients. This distinction alleviates concerns raised by Gorbatikov and Sikorskaya (2022), where observed concentration in the primary market may be an artifact of one AP representing the arbitrage activities of all its clients, rather than a reflection of a genuinely small number of arbitrageurs.

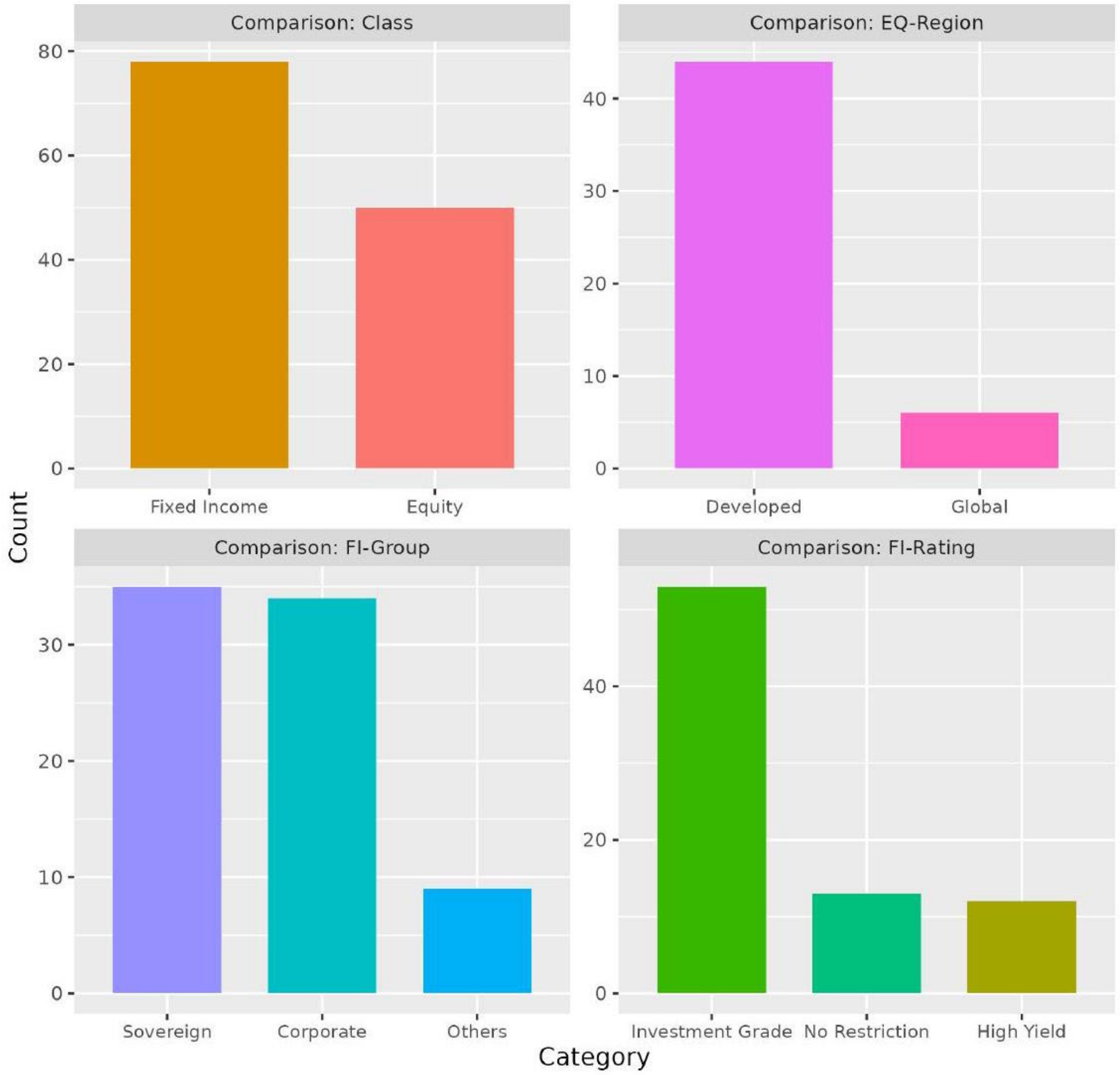
4.2 Secondary Market Data

We work with the reported trade data from the MDP dataset under the MiFID II reporting guidelines provided by the FCA. The transactions we analyse include: (i) all transactions where the ETF is listed or cross-listed in the UK; (ii) all transactions of ETFs traded on trading venues with reporting obligations to the FCA; (iii) all transactions of ETFs executed by firms with reporting obligations to the FCA. Due to its jurisdictional limitations, we start with ETFs that are primarily traded on the LSE to ensure our dataset has good coverage of the secondary market. We then filtered out some ISINs with a large number of outlier prices when compared to Bloomberg data[5].

To study the primary market activities of APs, we focus on ETFs from two of the largest issuers in our sample, as we have access to their primary market data. In the end, we have a sample of 128 ETFs, 50 of which are Equity ETFs and 78 of which are Fixed-income ETFs. The composition of their underlying types is presented in Figure 5.

[5] We compare the reported price with the price for the primary ticker of an ISIN in Bloomberg. Given that our data includes both on- and off-exchange trades reported by firms, and that each trade is denoted by the ISIN instead of the ticker, small deviations between the reported price and daily prices on Bloomberg are expected. To allow for this discrepancy and maintain trust in the validity of the reported data (as it is a legal obligation for firms to report accurately), we use a generous rule for defining outliers: if an ISIN has over 5% of currency-corrected reported prices outside the [2min-max, 2 max-min] range (where max and min refer to the maximum and minimum prices over the entire sample period provided by Bloomberg), we regard the ISIN as an outlier and exclude it from the analysis.

Figure 5: Sample ETFs by Categories



The MDP data offers very detailed information for each transaction. The fields most relevant to our study are timestamp (with microsecond frequency), ISIN, execution price, quantity, buyer, seller, buyer/seller decision makers, reporting and intermediary identities (and whether they are institutional or retail traders as identified by the Legal Entity Identifier (LEI)), and the trade capacity of the reporting entity. To make the transaction dataset suitable for our purposes, we conducted the following cleaning steps: (i) removed intra-trades as specified by the Level II dataset; (ii) mapped out the chains of transactions, removed duplicate reports by entities on the same chain, and (iii) deleted middle entities who were simply matching the two sides of a transaction without taking on risk. Each of these steps is discussed in detail in the Annex.

4.3 Other financial variables

We complement our dataset with ETF and AP-specific information from Bloomberg. For ETFs, this includes fund expense ratios, tracking errors, inception dates, current market cap, primary tickers for the ETFs, as well as time series of daily ETF closing prices, NAVs[6] average bid-ask spread as a percentage of mid-price, and total shares outstanding.

For macro control variables, we collect from Bloomberg the benchmark S&P 500 returns and the Bloomberg Global-Aggregate Total Return Index Value, which measures the performance of global investment-grade fixed-income markets. Additionally, we collect macroeconomic controls from the Federal Reserve Bank of St. Louis, including the VIX, the one-year constant maturity Treasury yield, the spread between the 10-Year and 2-Year Treasury constant maturity rates, and the ICE BofA US Corporate Index Option-Adjusted Spread, which is the calculated spread between a computed OAS index of US dollar-denominated investment-grade corporate debt and a spot Treasury curve.

4.4 Summary statistics

From Figure 5, we can see that the dataset we use covers a broad range of ETFs, split into Fixed Income and Equity categories. Our Fixed Income ETFs predominantly focus on sovereign and corporate bond ETFs, with a significant portion in investment-grade securities. In contrast, our Equity ETFs primarily target developed markets, with a smaller number providing exposure to global equities. This broad coverage allows us to investigate the heterogeneity in the relationship between APs' price and inventory decisions and ETFs' mispricing.

Looking at the summary statistics tables 1 and 2, expense ratios are fairly similar between the two categories, with both Fixed Income and Equity ETFs having median expense ratios around 0.2%. However, Fixed Income ETFs show slightly lower expense ratios at lower quantiles, which could be attributed to their simpler management and less frequent rebalancing requirements compared to equity funds. Comparing this with MEAN_MISPRICING, which is the mean mispricing across the 5 -year period for each ETF, we find that in general, they are on similar scales (around $0.1\% - 0.5\%$). This makes the magnitude of deviation during stress periods $(2\% - 5\%)$ more striking. A more pronounced difference between Fixed Income and Equity ETFs is seen in ETF tracking error, where Equity ETFs exhibit a much higher median value of 4.475%, compared to 1.691% for Fixed Income ETFs. This disparity likely stems from the inherent volatility of equity markets and the complexities involved in replicating indices that include global equities, which are subject to market timing issues and fluctuations across different time zones. In contrast, bond markets, particularly investment-grade bonds, are more stable, contributing to lower tracking error for Fixed Income ETFs. For both asset classes, NAV tracking errors are generally smaller than ETF tracking errors, potentially because ETFs are subject to intraday price fluctuations, liquidity dynamics, and transaction costs that introduce more volatility in their tracking accuracy whereas NAVs are much more stable. The NAV tracking error for Fixed Income can be much higher than for Equity, consistent

[6] NAV in Bloomberg includes a timing adjustment where estimated values are recorded for underlyings that are not traded during the same hours as the ETFs.

with the observation in Koont et al. (2022) that the creation/redemption basket may deviate from the index the ETF is tracking, particularly when the underlying is less liquid.

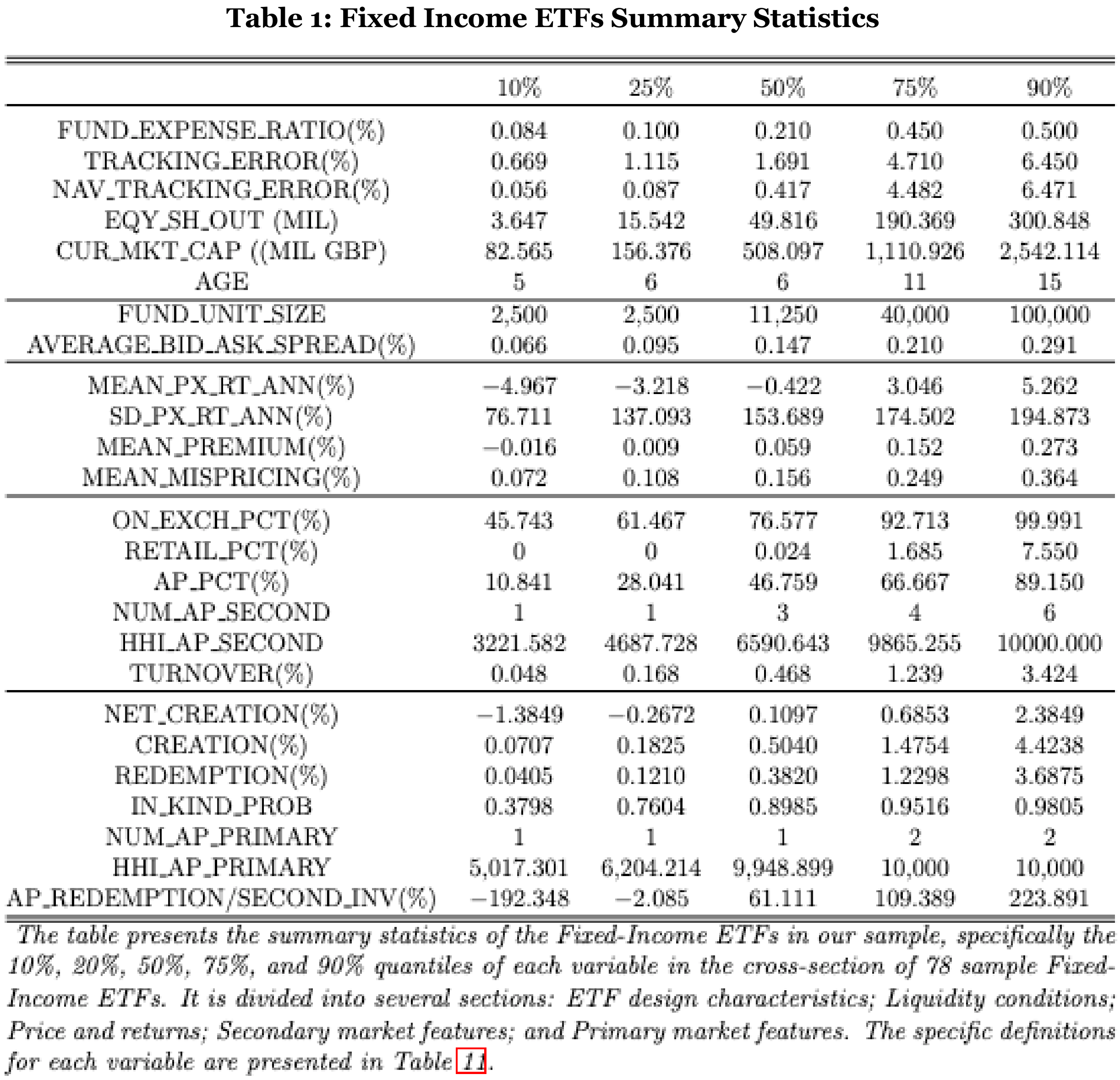
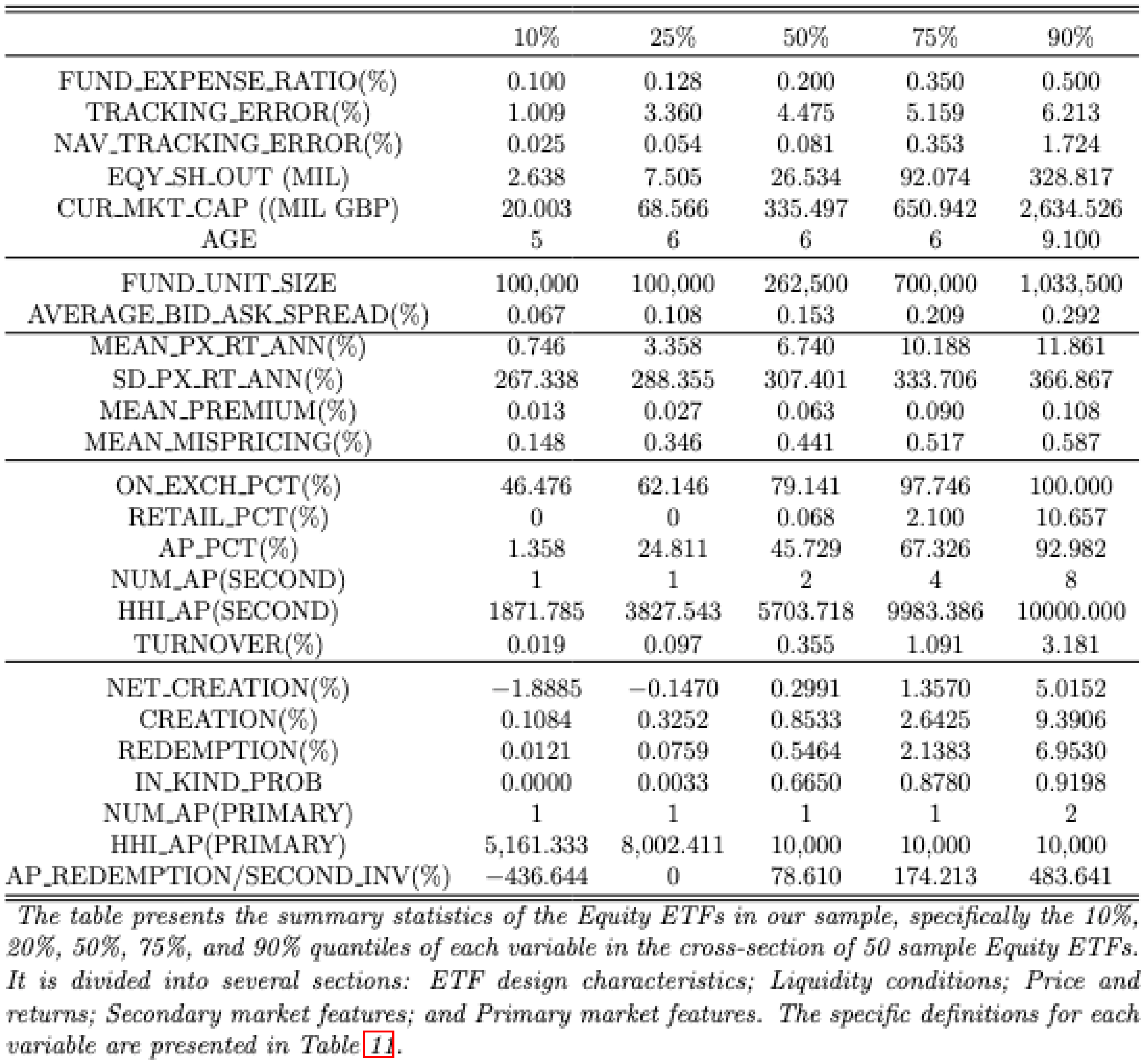


Table 1: Fixed Income ETFs Summary Statistics

	10%	25%	50%	75%	90%
FUND_EXPENSE_RATIO(%)	0.084	0.100	0.210	0.450	0.500
TRACKING_ERROR(%)	0.669	1.115	1.691	4.710	6.450
NAV_TRACKING_ERROR(%)	0.056	0.087	0.417	4.482	6.471
EQY_SH_OUT (MIL)	3.647	15.542	49.816	190.369	300.848
CUR_MKT_CAP ((MIL GBP)	82.565	156.376	508.097	1,110.926	2,542.114
AGE	5	6	6	11	15
FUND_UNIT_SIZE	2,500	2,500	11,250	40,000	100,000
AVERAGE_BID_ASK_SPREAD(%)	0.066	0.095	0.147	0.210	0.291
MEAN_PX_RT_ANN(%)	−4.967	−3.218	−0.422	3.046	5.262
SD_PX_RT_ANN(%)	76.711	137.093	153.689	174.502	194.873
MEAN_PREMIUM(%)	−0.016	0.009	0.059	0.152	0.273
MEAN_MISPRICING(%)	0.072	0.108	0.156	0.249	0.364
ON_EXCH_PCT(%)	45.743	61.467	76.577	92.713	99.991
RETAIL_PCT(%)	0	0	0.024	1.685	7.550
AP_PCT(%)	10.841	28.041	46.759	66.667	89.150
NUM_AP_SECOND	1	1	3	4	6
HHI_AP_SECOND	3221.582	4687.728	6590.643	9865.255	10000.000
TURNOVER(%)	0.048	0.168	0.468	1.239	3.424
NET_CREATION(%)	−1.3849	−0.2672	0.1097	0.6853	2.3849
CREATION(%)	0.0707	0.1825	0.5040	1.4754	4.4238
REDEMPTION(%)	0.0405	0.1210	0.3820	1.2298	3.6875
IN_KIND_PROB	0.3798	0.7604	0.8985	0.9516	0.9805
NUM_AP_PRIMARY	1	1	1	2	2
HHI_AP_PRIMARY	5,017.301	6,204.214	9,948.899	10,000	10,000
AP_REDEMPTION/SECOND_INV(%)	−192.348	−2.085	61.111	109.389	223.891

The table presents the summary statistics of the Fixed-Income ETFs in our sample, specifically the 10%, 20%, 50%, 75%, and 90% quantiles of each variable in the cross-section of 78 sample Fixed-Income ETFs. It is divided into several sections: ETF design characteristics; Liquidity conditions; Price and returns; Secondary market features; and Primary market features. The specific definitions for each variable are presented in Table 11.

Table 2: Equity ETFs Summary Statistics



	10%	25%	50%	75%	90%
FUND_EXPENSE_RATIO(%)	0.100	0.128	0.200	0.350	0.500
TRACKING_ERROR(%)	1.009	3.360	4.475	5.159	6.213
NAV_TRACKING_ERROR(%)	0.025	0.054	0.081	0.353	1.724
EQY_SH_OUT (MIL)	2.638	7.505	26.534	92.074	328.817
CUR_MKT_CAP ((MIL GBP)	20.003	68.566	335.497	650.942	2,634.526
AGE	5	6	6	6	9.100
FUND_UNIT_SIZE	100,000	100,000	262,500	700,000	1,033,500
AVERAGE_BID_ASK_SPREAD(%)	0.067	0.108	0.153	0.209	0.292
MEAN_PX_RT_ANN(%)	0.746	3.358	6.740	10.188	11.861
SD_PX_RT_ANN(%)	267.338	288.355	307.401	333.706	366.867
MEAN_PREMIUM(%)	0.013	0.027	0.063	0.090	0.108
MEAN_MISPRICING(%)	0.148	0.346	0.441	0.517	0.587
ON_EXCH_PCT(%)	46.476	62.146	79.141	97.746	100.000
RETAIL_PCT(%)	0	0	0.068	2.100	10.657
AP_PCT(%)	1.358	24.811	45.729	67.326	92.982
NUM_AP(SECOND)	1	1	2	4	8
HHI_AP(SECOND)	1871.785	3827.543	5703.718	9983.386	10000.000
TURNOVER(%)	0.019	0.097	0.355	1.091	3.181
NET_CREATION(%)	−1.8885	−0.1470	0.2991	1.3570	5.0152
CREATION(%)	0.1084	0.3252	0.8533	2.6425	9.3906
REDEMPTION(%)	0.0121	0.0759	0.5464	2.1383	6.9530
IN_KIND_PROB	0.0000	0.0033	0.6650	0.8780	0.9198
NUM_AP(PRIMARY)	1	1	1	1	2
HHI_AP(PRIMARY)	5,161.333	8,002.411	10,000	10,000	10,000
AP_REDEMPTION/SECOND_INV(%)	−436.644	0	78.610	174.213	483.641

The table presents the summary statistics of the Equity ETFs in our sample, specifically the 10%, 20%, 50%, 75%, and 90% quantiles of each variable in the cross-section of 50 sample Equity ETFs. It is divided into several sections: ETF design characteristics; Liquidity conditions; Price and returns; Secondary market features; and Primary market features. The specific definitions for each variable are presented in Table 11.

In terms of liquidity, both categories maintain tight bid-ask spreads, although the spreads widen at higher percentiles, indicating that more specialized funds, particularly those in less liquid markets, experience slightly higher trading costs. Creation unit sizes are much larger for Equity ETFs, with a median of 262,500 shares, compared to 11,250 for Fixed Income ETFs. This larger size suggests a higher level of institutional participation in equity markets, where liquidity is more abundant and transaction costs are lower. MEAN_MISPRICING is an important feature to highlight, with Equity ETFs showing greater deviations from their net asset value (NAV). The median mispricing for Equity ETFs stands at 0.441%, compared to 0.156% for Fixed Income ETFs. This higher mispricing in Equity ETFs is likely due to greater volatility and market dislocations, especially in global markets where differing trading hours and local conditions can result in temporary price inefficiencies. Fixed Income ETFs generally experience less mispricing because of the relative stability of bond markets and their more predictable price movements. However, during periods of market stress, especially in less liquid segments such as high-yield bonds, mispricing in Fixed Income ETFs can increase as liquidity constraints make it more difficult for market participants to arbitrage price discrepancies (see for example Figure 6).

The secondary market data reveals that both Fixed Income and Equity ETFs have most of their trading occurring on exchanges, although about 20% of the trading still happens off-exchange. Trades involving retail traders constitute a very small percentage of the total volume (within 10%). In contrast, APs have a significant presence in the secondary market for both Fixed Income and Equity ETFs. Trades involving an AP account for a median of around 46% of the daily turnover in both categories. The data further show that both categories have a highly concentrated AP market, as indicated by the high Herfindahl Hirschman Index (HHI) values and the small number of active APs (median value being $2-3$) on a given day, particularly in Fixed Income ETFs. This concentration suggests that a few large APs dominate the creation and redemption of ETF shares, which can impact pricing efficiency. This is consistent with concerns that in times of market stress, APs may be reluctant to create or redeem shares, especially in less liquid markets, which can exacerbate mispricing[7].

In terms of the primary market's creation and redemption activities, while the total shares outstanding and market cap of Fixed Income ETFs may be larger, Equity ETFs experience more frequent creation and redemption activities due to higher trading volumes, short-term strategies, greater use by a variety of investors, and the relatively higher volatility in equity markets. Fixed Income ETFs are also more likely to be created/redeemed in-kind. This distinctive feature of ETFs (compared with mutual funds) makes them less prone to runs on the fund during stress periods, as it shifts the cost of liquidation from the remaining investors in the fund to the redeeming investors. The fact that Fixed Income ETFs see more in-kind trading is intuitive, as they generally have higher costs of liquidation. The participation of APs in the primary market appears even more concentrated, with a median value of 1 AP. This suggests that there may be inter-AP trades in the secondary market, with larger APs aggregating orders from smaller APs before trading in the primary market.

Linking the AP's trades between the primary and secondary markets, we further find that AP_REDEMPTION/SECOND_INV (%) (defined as the proportion of shares that the AP net redeems (daily total redemption - daily total creation) as a percentage of the net position they accumulated from the secondary market on a given day) could be either negative or positive. This demonstrates that APs do not use the primary market solely to manage demand from the secondary market (which would typically result in the variable being within 0 and 1); they may also have arbitrage or inventory management objectives. This motivates our model of the AP's role as a market maker, arbitrageur, and inventory manager, with different weightings on the three objectives depending on the type of AP. For example, a bank AP might place more weight on inventory management compared to a high-frequency trader AP, who would prioritize arbitrage.

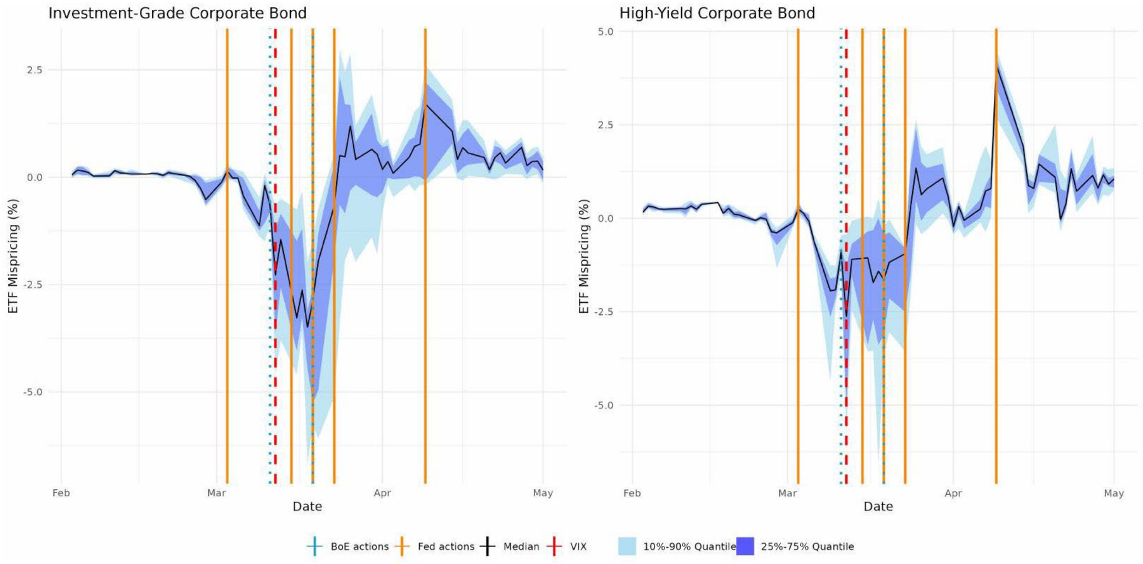
[7] Some empirical studies have documented the reassuring observation that alternative APs would step up to provide liquidity during times of stress Aquilina et al. (2021).

4.5 Covid Mispricing

Our sample covers the volatile March 2020 Covid period, and the significant mispricing observed across various asset classes during this time draws attention to the drivers behind it. We present in Figures 6, 7, 8, and 9, the mispricing for different types of ETFs in our dataset from February 15 to May 15, 2020. Figure 10 shows the mispricing of our representative ETF, zooming in on the same Covid period. The black line in each chart represents the median price premium, while the light blue and purple shaded areas show the dispersion across ETFs of the same type, from the 10%-90% and 25%-75% quantiles. These dispersions provide insight into how consistently ETFs within the same category were priced relative to their net asset values (NAVs), and how market stress and policy actions affected different funds.

In investment-grade corporate bond ETFs (See Figure 6), the price premium remains relatively stable early in the period, with narrow dispersion among the funds, as indicated by the small shaded areas. However, starting in mid-March, we observe a significant negative mispricing of more than 5%, and the dispersion widens significantly, particularly following the Fed's interventions. The expansion of the dispersion bands shows that not all ETFs within this category responded uniformly to evolving market conditions, despite the Fed's actions. Recovery started around March 19, when the Bank of England (BoE) had a second round of rate cuts for investment-grade corporate bond ETFs. Around March 23, when the Fed announced the PMCCF and SMCCF, we see a sharp recovery in high-yield corporate bond ETFs. Before April 9, there were some upward trends in mispricing for both categories as investor confidence in the ETF market improved, but the underlying corporate bond market had yet to catch up. The announcement of the expansion of both credit facilities restored liquidity in the underlying market and quickly reverted the trend. Interestingly, while high-yield bonds generally carry more credit risk, the dispersion across ETFs in this category during mid-March was narrower compared to investment-grade corporate bond ETFs. The median mispricing did turn significantly negative but to a smaller extent compared to investment-grade corporate bond ETFs. This suggests that the high-yield ETF market, though more volatile, responded more uniformly to market stress than the investment-grade sector, possibly due to more consistent risk pricing in the high-yield market or lower overall liquidity, which limits significant price deviations.

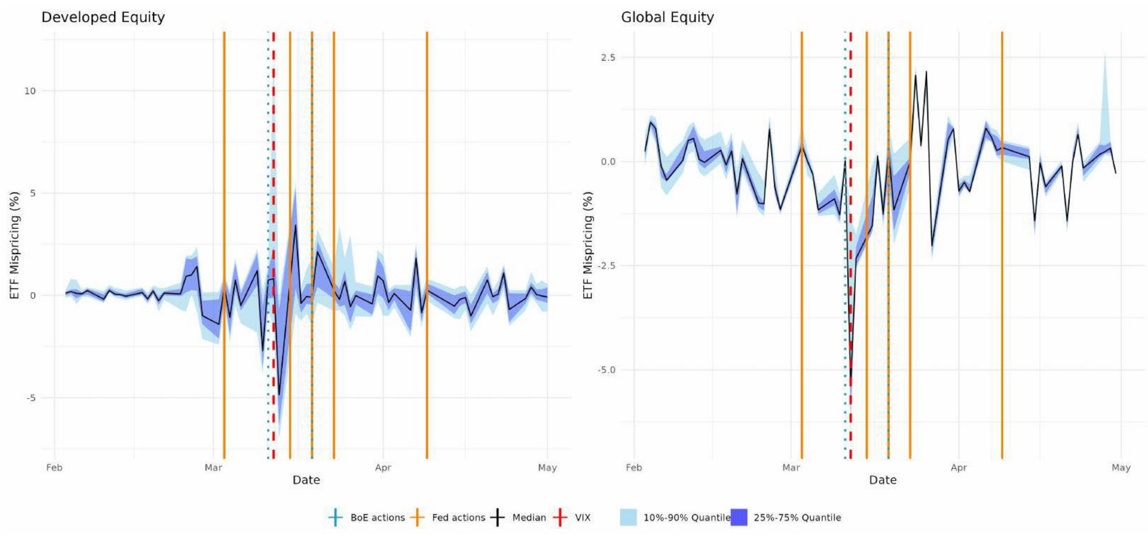
Figure 6: Corporate Bond ETF Premium During Covid



The above figure demonstrates the ETF premium, defined as $100 \times \frac{\text{Price-NAV}}{NAV}$*, for 20 Investment Grade and 12 High-Yield Corporate Bond ETFs in our sample between February 15, 2020, and May 15, 2020. The light blue shade represents the* 10% *to* 90% *quantile, the purple shade depicts the inter-quantile range, while the black line shows the median value across ETFs. The vertical lines in the plot highlight key event dates during this 3-month period. The red line indicates March 12, when the implied volatility index (VIX) reached its highest value since the Great Financial Crisis. The orange lines represent key Federal Reserve actions: from left to right, March 3 marks the Fed's emergency rate cut of 50 basis points, March 15 marks another emergency rate cut of 100 basis points along with an announcement to increase holdings of U.S. Treasury and mortgage-backed securities, March 23 marks the establishment of the Primary Market Corporate Credit Facility (PMCCF) and Secondary Market Corporate Credit Facility (SMCCF) (the SMCCF included the purchase of corporate bond ETFs to stabilize the financial market), and April 9 marks the expansion of the PMCCF and SMCCF to increase the amount of ETF purchases. The green lines represent Bank of England (BoE) actions: March 11 marks an emergency rate cut of 50 basis points, followed by an additional rate cut of 15 basis points on March 19. It is important to note that March 19 is also the first trading day after the Fed announced the Money Market Mutual Fund Liquidity Facility (MMLF), which aimed to enhance the liquidity and stability of money market mutual funds.*

The equity ETFs (See Figure 7) show a noticeable increase in volatility and dispersion around mid-March, following the spike in market volatility and central bank actions. Global equity ETFs exhibited a large negative mispricing before gradually recovering to normal levels, similar to corporate bond ETFs. Developed equity ETFs, in contrast, experienced significant negative mispricing before quickly rebounding to a large positive mispricing as market confidence recovered. A few of the developed equity ETFs even showed premiums of up to 8% around the time of the BoE's emergency 50 basis point rate cut, indicating a flight-to-safety driven increase in demand for those ETFs.

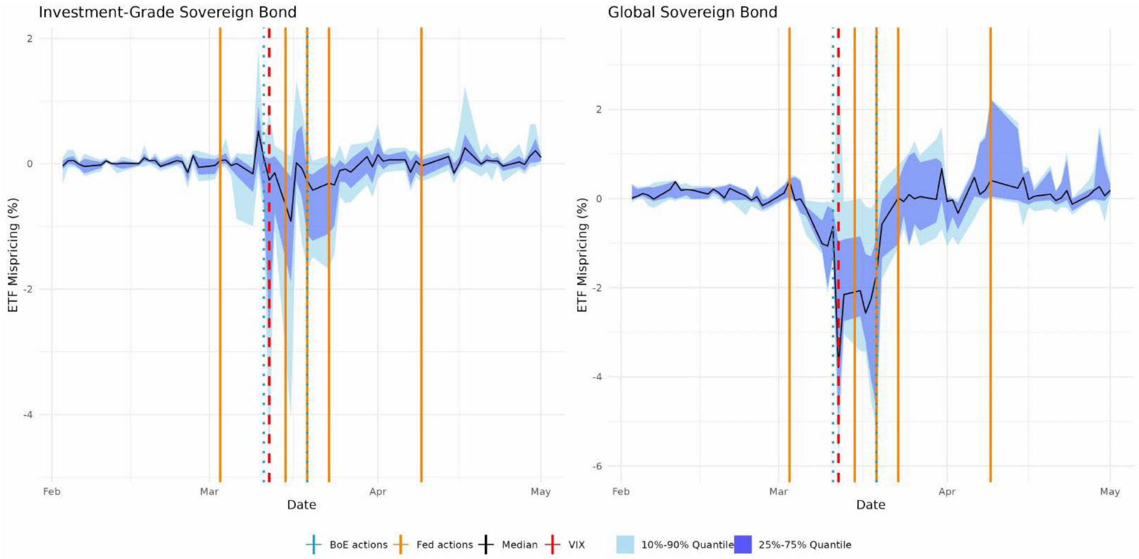
Figure 7: Equity ETF Premium During Covid



The above figure demonstrates the ETF premium, defined as $100 \times \frac{\textit{Price-NAV}}{NAV}$*, for 44 Developed and 6 Global Equity ETFs in our sample between February 15, 2020, and May 15, 2020. The light blue shade represents the* 10% *to* 90% *quantile, the purple shade depicts the inter-quantile range, while the black line shows the median value across ETFs. The vertical lines in the plot highlight key event dates during this 3-month period. The red line indicates March 12, when the implied volatility index (VIX) reached its highest value since the Great Financial Crisis. The orange lines represent key Federal Reserve actions: from left to right, March 3 marks the Fed's emergency rate cut of 50 basis points, March 15 marks another emergency rate cut of 100 basis points along with an announcement to increase holdings of U.S. Treasury and mortgage-backed securities, March 23 marks the establishment of the Primary Market Corporate Credit Facility (PMCCF) and Secondary Market Corporate Credit Facility (SMCCF) (the SMCCF included the purchase of corporate bond ETFs to stabilize the financial market), and April 9 marks the expansion of the PMCCF and SMCCF to increase the amount of ETF purchases. The green lines represent Bank of England (BoE) actions: March 11 marks an emergency rate cut of 50 basis points, followed by an additional rate cut of 15 basis points on March 19. It is important to note that March 19 is also the first trading day after the Fed announced the Money Market Mutual Fund Liquidity Facility (MMLF), which aimed to enhance the liquidity and stability of money market mutual funds.*

For sovereign bond ETFs (See Figure 8), the patterns differ somewhat. Investment-grade sovereign bond ETFs show relatively narrow dispersion throughout the period, with only a modest widening of the shaded areas in mid-March. This suggests that these ETFs were generally priced more consistently than corporate or equity ETFs, likely due to the perceived safety and liquidity of government bonds. The narrower bands indicate less variation in mispricing among these funds, even during periods of heightened market stress. On the other hand, global sovereign bond ETFs displayed slightly wider dispersion, and the recovery to usual levels happened around the time when both the BoE and the Fed made their second round of rate cuts.

Figure 8: Sovereign Bond ETF Premium During Covid

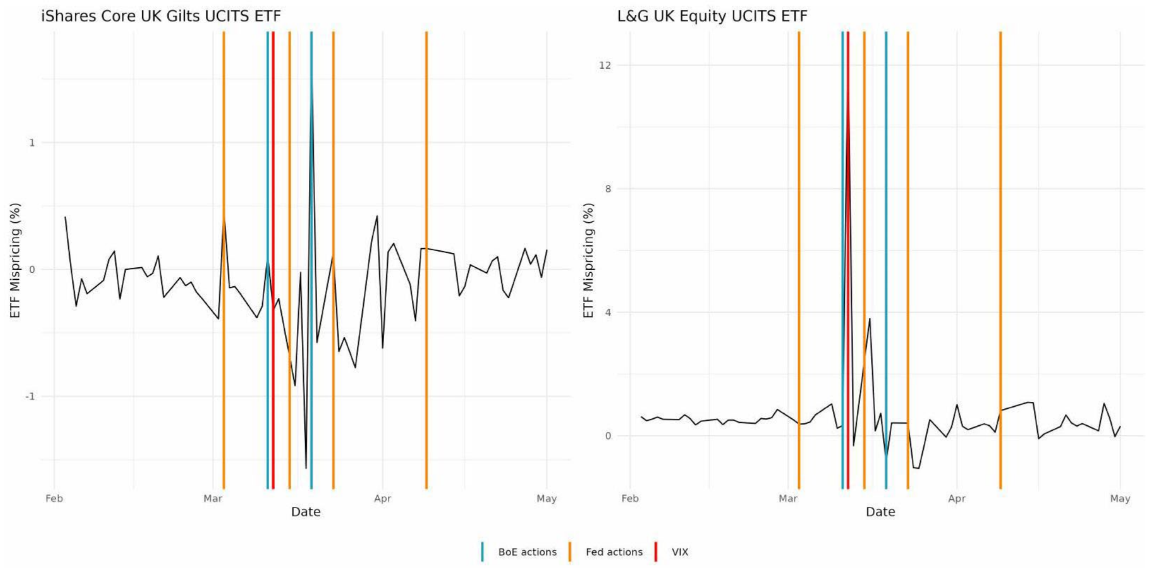


The above figure demonstrates the ETF premium, defined as $100 \times \frac{\textit{Price-NAV}}{\textit{NAV}}$*, for 24 Investment-Grade and 11 Global Sovereign Bond ETFs in our sample between February 15, 2020, and May 15, 2020. The light blue shade represents the* 10% *to* 90% *quantile, the purple shade depicts the inter-quantile range, while the black line shows the median value across ETFs. The vertical lines in the plot highlight key event dates during this 3-month period. The red line indicates March 12, when the implied volatility index (VIX) reached its highest value since the Great Financial Crisis. The orange lines represent key Federal Reserve actions: from left to right, March 3 marks the Fed's emergency rate cut of 50 basis points, March 15 marks another emergency rate cut of 100 basis points along with an announcement to increase holdings of U.S. Treasury and mortgage-backed securities, March 23 marks the establishment of the Primary Market Corporate Credit Facility (PMCCF) and Secondary Market Corporate Credit Facility (SMCCF) (the SMCCF included the purchase of corporate bond ETFs to stabilize the financial market), and April 9 marks the expansion of the PMCCF and SMCCF to increase the amount of ETF purchases. The green lines represent Bank of England (BoE) actions: March 11 marks an emergency rate cut of 50 basis points, followed by an additional rate cut of 15 basis points on March 19. It is important to note that March 19 is also the first trading day after the Fed announced the Money Market Mutual Fund Liquidity Facility (MMLF), which aimed to enhance the liquidity and stability of money market mutual funds.*

We also present in Figure 9 a UK gilt and a UK equity ETF. The Gilt ETF appears to have been much less affected by the stress period, with mispricing around 1% before quickly reverting after the BoE's additional 15 basis point rate cut. The UK equity ETF, on the other hand, mostly maintained a positive premium, which shot up to near 12% when the

VIX reached its highest value, indicating a temporary surge in demand for liquidity during times of stress.

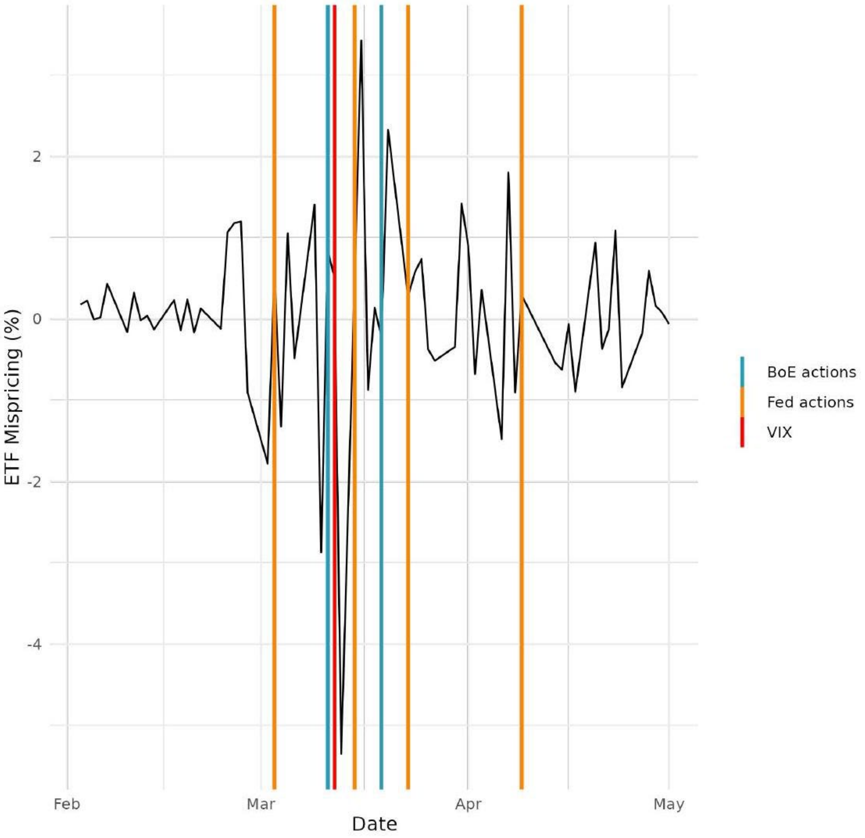
Figure 9: UK ETF Premium During Covid



The above figure demonstrates the ETF premium, defined as $100 \times \frac{\textit{Price-NAV}}{NAV}$*, for two major UK ETFs in our sample between February 15, 2020, and May 15, 2020. The light blue shade represents the* 10% *to* 90% *quantile, the purple shade depicts the inter-quantile range, while the black line shows the median value across ETFs. The vertical lines in the plot highlight key event dates during this 3-month period. The red line indicates March 12, when the implied volatility index (VIX) reached its highest value since the Great Financial Crisis. The orange lines represent key Federal Reserve actions: from left to right, March 3 marks the Fed's emergency rate cut of 50 basis points, March 15 marks another emergency rate cut of 100 basis points along with an announcement to increase holdings of U.S. Treasury and mortgage-backed securities, March 23 marks the establishment of the Primary Market Corporate Credit Facility (PMCCF) and Secondary Market Corporate Credit Facility (SMCCF) (the SMCCF included the purchase of corporate bond ETFs to stabilize the financial market), and April 9 marks the expansion of the PMCCF and SMCCF to increase the amount of ETF purchases. The green lines represent Bank of England (BoE) actions: March 11 marks an emergency rate cut of 50 basis points, followed by an additional rate cut of 15 basis points on March 19. It is important to note that March 19 is also the first trading day after the Fed announced the Money Market Mutual Fund Liquidity Facility (MMLF), which aimed to enhance the liquidity and stability of money market mutual funds.*

Figure 10 zoomed in on the Covid period for our representative ETF discussed in Figure 1.

Figure 10: iShares Core MSCI World UCITS ETF Premium During Covid



The above figure demonstrates the ETF premium, defined as $100 \times \frac{\text{price-NAV}}{\text{NAV}}$*, for the iShares Core MSCI World UCITS ETF between February 15, 2020, and May 15, 2020. The light blue shade represents the* 10% *to* 90% *quantile, the purple shade depicts the inter-quantile range, while the black line shows the median value across ETFs. The vertical lines in the plot highlight key event dates during this 3-month period. The red line indicates March 12, when the implied volatility index (VIX) reached its highest value since the Great Financial Crisis. The orange lines represent key Federal Reserve actions: from left to right, March 3 marks the Fed's emergency rate cut of 50 basis points, March 15 marks another emergency rate cut of 100 basis points along with an announcement to increase holdings of U.S. Treasury and mortgage-backed securities, March 23 marks the establishment of the Primary Market Corporate Credit Facility (PMCCF) and Secondary Market Corporate Credit Facility (SMCCF) (the SMCCF included the purchase of corporate bond ETFs to stabilize the financial market), and April 9 marks the expansion of the PMCCF and SMCCF to increase the amount of ETF purchases. The green lines represent Bank of England (BoE) actions: March 11 marks an emergency rate cut of 50 basis points, followed by an additional rate cut of 15 basis points on March 19. It is important to note that March 19 is also the first trading day after the Fed announced the Money Market Mutual Fund Liquidity Facility (MMLF), which aimed to enhance the liquidity and stability of money market mutual funds.*

In summary, the patterns in the charts reveal how price premiums across ETFs of the same category diverged during periods of market stress, with key policy actions, such as the Fed's rate cuts and the launch of the PMCCF and SMCCF, attempting to stabilize the markets. These interventions affected mispricing through multiple channels. In terms of restoring functional arbitrage for the APs, the rate cuts reduced the inventory cost of holding ETFs and acting as buyers of ETFs and the underlying securities mitigated both excess selling and the price impact of APs selling in the underlying market. Meanwhile, these interventions injected liquidity to revive less liquid underlying markets, allowing trading to resume and the NAV to become more reflective of market value. Some concerns were raised about instruments such as the SMCCF and MMLF diverting capital from ETF markets to others (e.g. Aramonte and Avalos (2020)), potentially worsening liquidity. However, such patterns were not observed in our sample.

Linking our derived mispricing dynamics (see Equation 70) to the Covid mispricing charts, we observe that ETF mispricing is persistent, responds directionally to current and past demand shocks, and reflects expectations about fundamental value - which may be influenced by government policies. However, APs' inventory is likely an important omitted variable in this analysis due to its significant impact on the mispricing dynamics. We therefore empirically examine this factor in Section 6.

5 Empirical Methods

5.1 Extraction of fundamental values

The gap between ETF price and NAV with the fundamental values is an important quantity that guides ETF arbitrage activities. Based on the dynamics derived in our model, we can formulate a state-space model to extract the fundamental value using only price and NAV information, as both the price and NAV are noisy measurements of it.

State-space Model

From Equation 47 and 67, we have

$$P = \mu + \alpha_1(I - I^d) + \alpha_2 x + \alpha_3(P^u - \mu), \tag{80}$$

$$P^{u'} = P^u - 2\theta_1(I - I^d) + \theta_2(P^u - \mu) + \theta_1 x = (1 + \theta_2)P^u - \theta_2\mu - 2\theta_1(I - I^d) + \theta_1 x. \tag{81}$$

Using the inventory dynamic (Equation 61) and the dynamic for $P^u - \mu$ (Equation 67), we have

$$\begin{bmatrix} I' - I^d \\ P^{u'} - \mu' \end{bmatrix} = \begin{bmatrix} 1 - \iota_1 & -\iota_3 \\ -2\theta_1 & 1 + \theta_2 \end{bmatrix} \begin{bmatrix} I - I^d \\ P^u - \mu \end{bmatrix} + \begin{bmatrix} -\iota_2 \\ \theta_1 \end{bmatrix} x + \begin{bmatrix} 0 \\ -\dfrac{\Omega}{(1 - \Omega)\delta} \end{bmatrix} x'. \tag{82}$$

This shows that any nonzero linear combination of $I - I^d$ and $P^u - \mu$ follows an ARMA $(1,1)$ process. In particular, for $N = \alpha_1(I - I^d) + \alpha_2 x + \alpha_3(P^u - \mu)$, we have

$$\begin{aligned} N' &= \mathrm{M}N + (-\alpha_1\iota_2 + \alpha_3\theta_1 - \mathrm{M}\alpha_2)x + (\alpha_2 - \alpha_3\frac{\Omega}{(1 - \Omega)\delta})x' \\ &= \psi_p N + e'_p\left(1 + \theta_p L\right) \end{aligned} \tag{83}$$

where $M = \frac{(1-\iota_1)\alpha_1^2 + (1+\theta_2)\alpha_3^2 - (\iota_3 + 2\theta_1)\alpha_1\alpha_3}{\alpha_1^2 + \alpha_3^2}$.

It can be verified that since $-\alpha_1\iota_2 + \alpha_3\theta_1 + \alpha_2 > 0$, there exists $\theta_p \in (-1,0). \psi_p \in (0,1)$ for most base parameter combinations, but it can exceed 1 when the inventory cost parameter, ϕw, is close to 0 (we don't consider this case here).

For $T = -2\theta_1(I - I^d) + \theta_1 x$, we have

$$\begin{aligned} T' &= \psi_T T + (2\iota_2 - \psi_T)\theta_1 x + \ \theta_1\, x' \\ &= \psi_T T + e'_n(1 + \theta_\mathrm{n}\, L) \end{aligned} \tag{84}$$

where $\psi_T = 1 - \iota_1$. Since $\iota_1 \in (0,1), \psi_T \in (0,1)$.
Therefore, we have

$$P^{u'} = \psi_n P^u + (1 - \psi_n)\mu + T, \tag{85}$$

where $\psi_n \in (0,1)$ as $\theta_2 \in (-1,0)$. Also, note that the dynamic for μ is

$$\mu' = (1-\Omega)\mu + \Omega v' + \frac{\Omega}{\delta}X'. \tag{86}$$

Therefore, we have

$$\mu = \frac{\Omega\eta}{(1-(1-\Omega)L)(1-L)} + \frac{\Omega X}{\delta(1-(1-\Omega)L)}, \tag{87}$$

and μ follows an $\mathrm{ARIMA}(1,1,1)$ with

$$\begin{aligned}\Delta\mu' &= (1-\Omega)\Delta\mu + \Omega\eta' + \frac{\Omega}{\delta}X'(1-L)\\ &= \psi_\mu\Delta\mu + e'_\mu\left(1+\theta_\mu L\right).\end{aligned} \tag{88}$$

Since $\Omega \in (0,1), \psi_\mu \in (0,1)$. We can also verify that $\theta_\mu \in (-1,0)$.
We then proceed with the estimation using the following simplified dynamic for price, NAV, and μ. For clarity, we use n to represent $P^{u\prime}$ (end-of-day NAV) and p to represent P (end-of-day price)

$$p = \mu + N = \frac{e_\mu\left(1+\theta_\mu L\right)}{(1-L)\left(1-\psi_\mu L\right)} + \frac{e_p\left(1+\theta_p L\right)}{\left(1-\psi_p L\right)}, \tag{89}$$

$$n = S + \tilde{T} = S + \frac{T}{1-\psi_n L} = (1-\psi_n)\frac{e_\mu\left(1+\theta_\mu L\right)}{(1-L)\left(1-\psi_\mu L\right)(1-\psi_n L)} + \frac{e_n(1+\theta_\mathrm{n}\, L)}{(1-\psi_n L)(1-\psi_T L)}, \tag{90}$$

where $S = (1-\psi_n)\frac{\mu}{1-\psi_n L}$. We can also rewrite the system into a state-space form as follows (where time subscripts are reintroduced)

$$\begin{bmatrix} p_t \\ n_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_t \\ \mu_{t-1} \\ S_t \\ S_{t-1} \\ S_{t-2} \\ N_t \\ \tilde{T}_t \\ \tilde{T}_{t-1} \\ \theta_\mu e_{\mu t} \\ \theta_p e_{pt} \\ \theta_n e_{nt} \end{bmatrix}, \tag{91}$$

$$\begin{bmatrix} \mu_t \\ \mu_{t-1} \\ S_t \\ S_{t-1} \\ S_{t-2} \\ N_t \\ \tilde{T}_t \\ \tilde{T}_{t-1} \\ \theta_\mu e_{\mu t} \\ \theta_p e_{pt} \\ \theta_n e_{nt} \end{bmatrix} = \begin{bmatrix} 1+\psi_\mu & -\psi_\mu & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1+\psi_n+\psi_\mu & -(\psi_n+\psi_\mu+\psi_n\psi_\mu) & \psi_n\psi_\mu & 0 & 0 & 0 & 1-\psi_n & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \psi_p & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \psi_n+\psi_T & -\psi_n\psi_T & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \mu_{t-2} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \\ N_{t-1} \\ \tilde{T}_{t-1} \\ \tilde{T}_{t-2} \\ \theta_\mu e_{\mu t-1} \\ \theta_p e_{pt-1} \\ \theta_n e_{nt-1} \end{bmatrix} \quad (92)$$

$$+ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1-\psi_n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \theta_\mu & 0 & 0 \\ 0 & \theta_p & 0 \\ 0 & 0 & \theta_n \end{bmatrix} \begin{bmatrix} e_{\mu t} \\ e_{pt} \\ e_{nt} \end{bmatrix},$$

We abstract the system as follows

$$y_t = Z\alpha_t,$$
$$\alpha_t = T\alpha_{t-1} + \eta_t, \eta_t \sim \mathcal{N}(0, \Omega_t),$$

where $\Omega_t = S_\eta Q_t S_\eta'$ and $Q_t = D_t R_t D_t . S_\eta$ is the selection matrix, D_t contains standard deviations, and R_t is the correlation matrix.

Estimation

As can be seen above, Δp_t is an $\text{ARMA}(2,3)$ and Δn_t is an ARMA (3,3). The 11 AR and MA coefficients from the two models and the 3 parameters from the covariance matrix can be used to identify $\psi_\mu, \psi_n, \psi_p, \psi_T, \theta_\mu, \theta_p$, and the 6 parameters in the covariance matrix for

$$[e_{\mu t} \quad e_{pt} \quad e_{nt}]'.$$

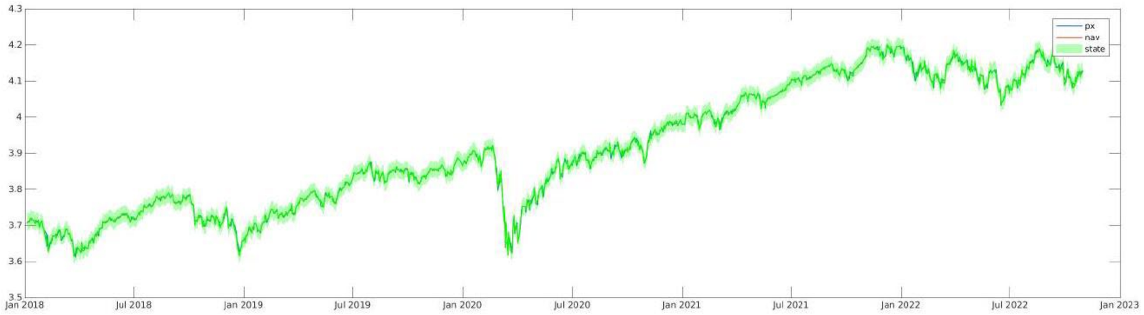
We initialize the MLE estimation by choosing a reasonable value for each parameter in the following way. We first calculate a ballpark estimate of the fundamental value by averaging the price and NAV. From Equation 88, fitting an $\text{ARMA}(1,1)$ to $\Delta\mu$ gives us initial values for ψ_μ, θ_μ, and σ_μ^2. From Equation 83, fitting an $\text{ARMA}(1,1)$ to $p - \mu$ provides initial values for ψ_p, θ_p, and σ_p^2. Lastly, from Equation 90, fitting an $\text{ARMA}(2,2)$ to Δn while taking $\Delta\mu_t$ and $\Delta\mu_{t-1}$ as exogenous variables deliver ψ_n, ψ_T, and σ_n^2, because we have

$$(1 - \psi_n L)(1 - \psi_T L)\Delta n_t = (1 - \psi_n)\Delta\mu_t - (1 - \psi_n)\psi_T \Delta\mu_{t-1} + e_n(1 + \theta_{\mathrm{n}} L)(1 - L).$$

The initial values for the correlation parameters are set to 0.

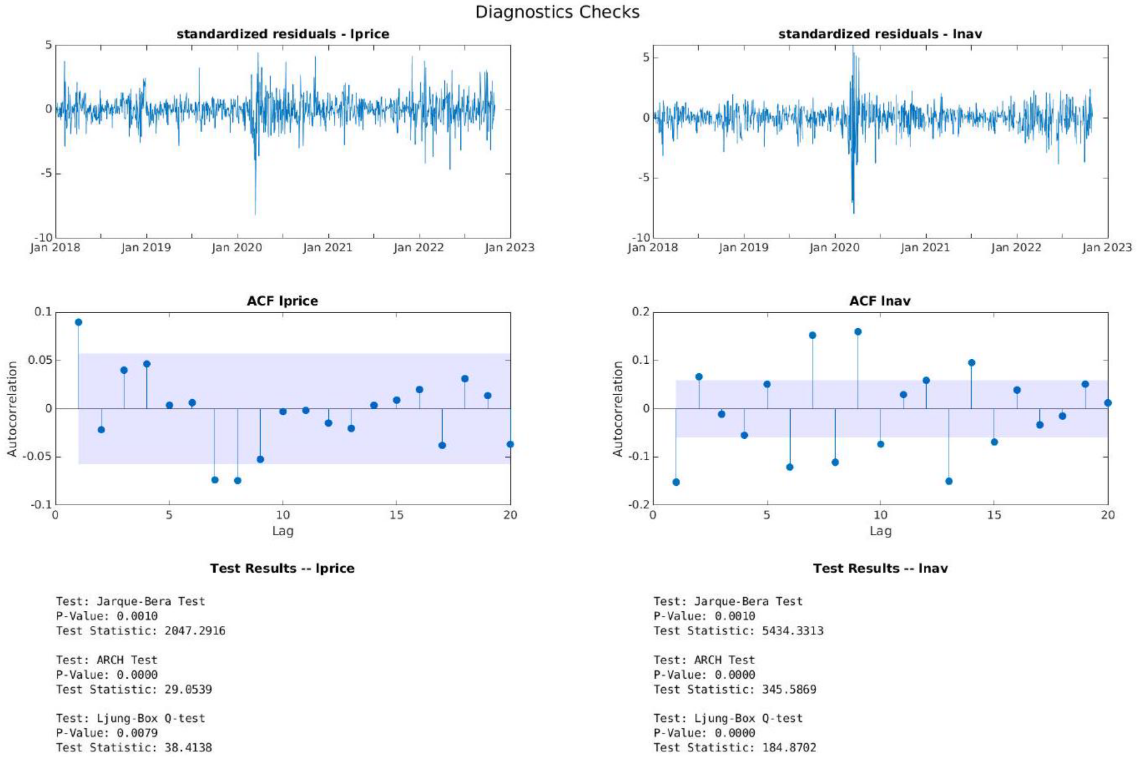
Baseline model results In Figure 11a and Figure 11b, we present the fitted Price, NAV, and fundamental value chart for our representative iShares Core MSCI World UCITS ETF (results for an alternative representative ETF are shown in Figure 21 in the Annex). The baseline model has a good fit based on the line chart. However, the diagnostics figure shows that there are still significant leftover dynamics in the residuals of our measurement equations that the state-space model is not capturing.

Figure 11: ETF Baseline Fitted Model for iShares Core MSCI World UCITS



(a) Price, NAV, Fundamental Value Chart

The above figure demonstrates the ETF price (blue), NAV (red), and the filtered first state variable (bright green) that corresponds to M_t or μ_t. The bands around the green line denote the 95% confidence interval constructed using the state forecast variance for each time period.



Test Results -- lprice

```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 2047.2916

Test: ARCH Test
P-Value: 0.0000
Test Statistic: 29.0539

Test: Ljung-Box Q-test
P-Value: 0.0079
Test Statistic: 38.4138
```

Test Results -- lnav

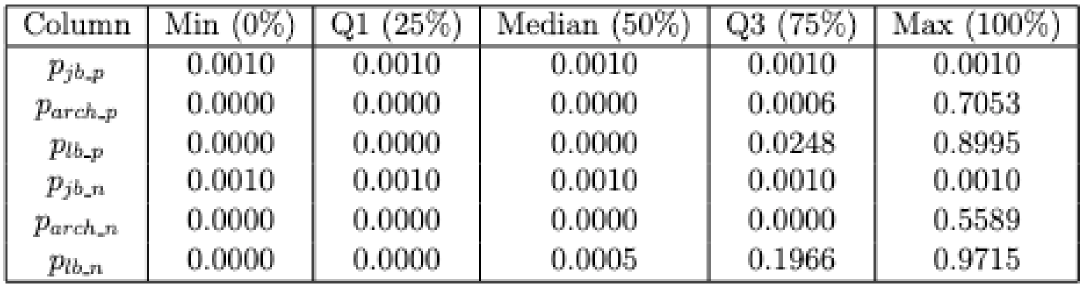
```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 5434.3313

Test: ARCH Test
P-Value: 0.0000
Test Statistic: 345.5869

Test: Ljung-Box Q-test
P-Value: 0.0000
Test Statistic: 184.8702
```

(b) Diagnostics

The above demonstrates the diagnostic checks for the baseline state-space model. The first row shows the standardized forecast residuals for predicting the two variables in the measurement equation: log (price) and log (NAV). The second row shows the ACF plots for the standardized residuals along with Bartlett's confidence interval. The third row presents the test statistics and p-values for the two measurement residuals: normality test, heteroskedasticity test, and autocorrelation test.

All the diagnostics tests are violated, including the Jarque-Bera normality test, the ARCH test for heteroskedasticity, and the Ljung-Box Q-test for serial autocorrelation. More generally, when we investigate the test results for all the ETFs in our sample, we observe from Table 3 that most of them fail to pass any diagnostic checks. We can further investigate the pattern of heteroskedasticity using non-parametric estimation.

Table 3: Summary Diagnostics for Sample ETFs from the baseline model



Column	Min (0%)	Q1 (25%)	Median (50%)	Q3 (75%)	Max (100%)
p_{jb_p}	0.0010	0.0010	0.0010	0.0010	0.0010
p_{arch_p}	0.0000	0.0000	0.0000	0.0006	0.7053
p_{lb_p}	0.0000	0.0000	0.0000	0.0248	0.8995
p_{jb_n}	0.0010	0.0010	0.0010	0.0010	0.0010
p_{arch_n}	0.0000	0.0000	0.0000	0.0000	0.5589
p_{lb_n}	0.0000	0.0000	0.0005	0.1966	0.9715

Non-parametric estimation of time-varying covariance matrix

Since we observe strong heteroskedasticity in the diagnostic checks for the error terms in the previous paragraph (especially during the Covid turmoil in March 2020), we can attempt to model this time-varying structure using nonparametric kernel smoothing and investigate further. Potential heteroskedasticity should not affect the consistency of the coefficients for the mean process. Therefore, we can retain the coefficient estimates from the previous paragraph and focus on the error process itself. Specifically, we can model the covariance matrix parameter as smooth functions

$$\sigma_{pt}^2 = \sigma_p^2\left(\frac{t}{T}\right), \sigma_{\mu t}^2 = \sigma_\mu^2\left(\frac{t}{T}\right), \sigma_{nt}^2 = \sigma_n^2\left(\frac{t}{T}\right),$$

and

$$\sigma_{p\mu t} = \sigma_{p\mu}\left(\frac{t}{T}\right), \sigma_{n\mu t} = \sigma_{n\mu}\left(\frac{t}{T}\right), \sigma_{pnt} = \sigma_{pn}\left(\frac{t}{T}\right).$$

We can rewrite the dynamics for Δp_t and Δn_t as:

$$\begin{bmatrix} \tilde{p}_t \\ \tilde{n}_t \end{bmatrix} = \begin{bmatrix} (1-\psi_\mu L)(1-\psi_p L)\Delta p_t \\ (1-\psi_\mu L)(1-\psi_n L)(1-\psi_T L)\Delta n_t \end{bmatrix},$$

and this is equivalently represented by:

$$\begin{bmatrix} \tilde{p}_t \\ \tilde{n}_t \end{bmatrix} = \begin{bmatrix} e_\mu(1+\theta_\mu L)(1-\psi_p L) + e_p(1+\theta_p L)(1-L)(1-\psi_\mu L) \\ (1-\psi_n)e_\mu(1+\theta_\mu L)(1-\psi_T L) + e_n(1+\theta_n L)(1-L)(1-\psi_\mu L) \end{bmatrix}. \quad (93)$$

With the estimates for $\psi_\mu, \psi_n, \psi_p, \psi_T, \theta_\mu$, θ_n and θ_p from the previous paragraph, we can write the covariance and first-order autocovariance of the LHS of Equation 93 as functions of the six covariance parameters. We can approximate the covariance as

$$\begin{bmatrix} c_1\sigma_{pt}^2 + c_2\sigma_{\mu t}^2 + c_3\sigma_{p\mu t} & c_4\sigma_{pnt} + c_5\sigma_{n\mu t} + c_6\sigma_{p\mu t} + c_7\sigma_{\mu t}^2 \\ c_4\sigma_{pnt} + c_5\sigma_{n\mu t} + c_6\sigma_{p\mu t} + c_7\sigma_{\mu t}^2 & c_8\sigma_{nt}^2 + c_9\sigma_{n\mu t} + c_{10}\sigma_{\mu t}^2 \end{bmatrix}, \tag{94}$$

with three unique terms. The first-order autocovariance is approximated as

$$\begin{bmatrix} c_{11}\sigma_{pt}^2 + c_{12}\sigma_{p\mu t} + c_{13}\sigma_{\mu t}^2 & c_{14}\sigma_{pnt} + c_{15}\sigma_{n\mu t} + c_{16}\sigma_{p\mu t} + c_{17}\sigma_{\mu t}^2 \\ c_{18}\sigma_{pnt} + c_{19}\sigma_{n\mu t} + c_{20}\sigma_{p\mu t} + c_{21}\sigma_{\mu t}^2 & c_{22}\sigma_{nt}^2 + c_{23}\sigma_{n\mu t} + c_{24}\sigma_{\mu t}^2 \end{bmatrix}, \tag{95}$$

with four unique terms. We can estimate the seven unique terms from the above time-varying covariance matrices using kernel smoothing. For example,

$$E[\tilde{p}_t\tilde{p}_{t-1}] = \frac{1}{T}\sum_{k=1}^{T} K_h\left(\frac{t-k}{T}\right)\tilde{p}_k\tilde{p}_{k-1}. \tag{96}$$

The bandwidth h is chosen according to Silverman's rule of thumb

$$h = T\frac{1.06\min\left\{\sigma_{t/T}, \frac{IQR(t/T)}{1.34}\right\}}{T^{0.2}}.$$

The six time-varying parameters can be recovered by minimizing the squared residuals from the seven equations[8] (all linear functions of the parameters), subject to the constraints that the variance terms are nonnegative, and the covariance terms imply a correlation bounded between -1 and 1. These constraints are formulated as nonlinear relationships between the covariance and variance parameters.

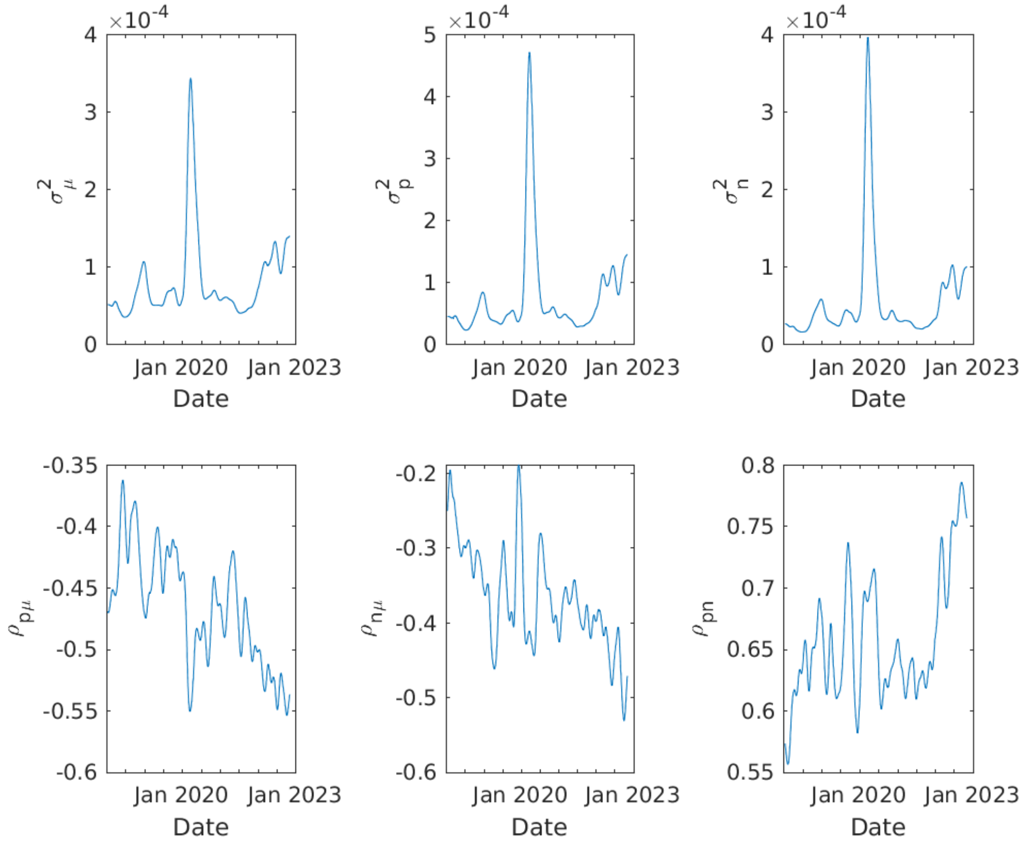
During the empirical estimation, we observed that in certain periods, some of the variance terms were estimated to be close to zero. This results in correlations approaching 1 or 1, which is unlikely. To improve the estimation, we exclude points where the estimated correlations exceed 0.99 or fall below -0.99.

The estimated time-varying mean variances and correlations across all sample ETFs[9] are presented in Figure 12. A significant jump in all three variances is observed in March 2020, where the variances increase to more than four times their normal levels. The uncertainty brought about by Covid led to a substantial increase in market noise and heightened uncertainty surrounding the fundamental value of ETFs. Among the three variances, the error term in the ETF price dynamic shows the greatest variability.

[8] In theory, since the RHS of Equation 93 is a VMA(3), we can use all three autocovariance matrices. However, incorporating all three does not significantly change the results, so we opt for a simpler approach here.

[9] The shown results in the plot has been smoothed again with the Silverman's bandwidth (roughly 80-day windows) for clearer presentation.

Figure 12: Mean Time Varying Error Variances and Correlations across sample ETFs (Nonparametric model)



Regarding the time-varying correlations, both $\rho_{p\mu}$ and $\rho_{n\mu}$ fluctuate between -0.4 and -0.7, with no significant deviations during the Covid period. The correlation between p and n is estimated to be around 0.8, but there appears to be a decrease in this high correlation during the Covid period, likely due to investors using ETFs and the underlying assets as hedges against each other.

Extension to score-driven model

As discussed above, there is significant heteroskedasticity in the data. Although the coefficients from the mean process remain consistent, they are less efficient. More importantly, heteroskedasticity can impact the extraction of the state variables, as their updates depend on the estimated error covariance matrix for each period.

The nonparametric method in the previous paragraph helps visualize the uncaptured dynamic covariance structure in the error terms with minimal assumptions. However, we also aim to use the time-varying covariance structure to improve the extraction of the state variables. Therefore, we extend our baseline state-space model to a score-driven state-space model, where the score drives the dynamics of the time-varying covariance parameters. This extension has the additional benefit of potentially providing a smoother pattern for the extracted covariance parameters, though it imposes the normality assumption.

In this model, we allow both the standard deviations and correlations to be time-varying. We stack the time-varying parameters into a vector f_t as follows

$$f_t = \begin{bmatrix}\delta_t \\ \gamma_t\end{bmatrix}, \delta_t = \begin{bmatrix}\log\,\sigma_{\mu t} \\ \log\,\sigma_{pt} \\ \log\,\sigma_{nt}\end{bmatrix}, \gamma_t = \begin{bmatrix}\text{atanh}\pi_{p\mu t} \\ \text{atanh}\pi_{n\mu t} \\ \text{atanh}\pi_{pnt}\end{bmatrix}$$

where π represents the partial correlations between pairs of error terms. Following Monache et al. (2021), we reparametrize correlations as partial correlations with the following correspondence

$$\rho_{p\mu t} = \pi_{p\mu t}, \rho_{n\mu t} = \pi_{n\mu t}, \rho_{pnt} = \pi_{pnt}\sqrt{\left(1-\pi_{p\mu t}^2\right)\left(1-\pi_{n\mu t}^2\right)} + \pi_{p\mu t}\pi_{n\mu t}.$$

This allows us to express the dynamic of time-varying coefficients as

$$f_{t+1} = \mathrm{c} + \mathrm{A}f_t + \mathrm{B}s_t, s_t = \mathcal{S}_t\nabla_t, t = 1, \dots, T, \tag{97}$$

with:

$$\nabla_t = \frac{\partial \ell_t}{\partial f_t}, \mathcal{S}_t = -\mathrm{E}_t\left(\frac{\partial^2 \ell_t}{\partial f_t \partial f_t'}\right)^{-1}.$$

where ∇_t is the score of the conditional log-likelihood function with respect to f_t, and $\mathcal{S}_t$ is the inverse of the information matrix. Therefore, s_t has a conditional mean of 0 and a conditional variance that is the inverse of the information matrix.

We define the following variables used in the state-space recursions[10]

$$v_t = y_t - Z\alpha_t, F_t = ZP_tZ',$$

$$a_{t|t} = a_t + P_tZ'F_t^{-1}v_t, P_{t|t} = P_t - P_tZ'F_t^{-1}ZP_t,$$
$$a_{t+1} = Ta_{t|t}, P_{t+1} = TP_{t|t}T' + \Omega_{t+1}.$$

Following Monache et al. (2021), we can derive the score and information matrix as follows (where $\dot{X}_t = \partial\text{vec}(X_t)/\partial f_t'$ for any matrix X_t)

$$\nabla_t = \frac{1}{2}\left[\dot{F}_t'(F_t \otimes F_t)^{-1}\text{vec}(v_tv_t' - F_t)\right], \tag{98}$$

[10] When there are missing values in y_t, we use the selection matrix W_t and redefine the observation equation as follows

$$W_ty_t = W_tZ\alpha_t.$$

The recursion is then modified as

$$v_t = W_t(y_t - Za_t), F_t = W_t(ZP_tZ')W_t',$$
$$a_{t|t} = a_t + P_tZ'W_t'F_t^{-1}v_t, P_{t|t} = P_t - P_tZ'W_t'F_t^{-1}W_tZP_{t,}$$

and $\dot{F}_t$ is modified as

$$\dot{F}_t = (W_tZ \otimes W_tZ)\dot{Q}_{t.}$$

Everything else in the model remains the same.

$$I_t = \frac{1}{2}[\dot{F}_t'(F_t \otimes F_t)^{-1}\dot{F}_t], \tag{99}$$

where

$$\dot{F}_t = (Z \otimes Z)\dot{\Omega}_t,$$
$$\dot{\Omega}_t = \left(S_\eta \otimes S_\eta\right)[(D_t R_t \otimes I + I \otimes D_t R_t)\dot{D}_t + (D_t \otimes D_t)\dot{R}_t],$$
$$\mathrm{vec}(D_t) = S_{1,d}\psi_d\left(S_{2,d}f_t\right), \mathrm{vec}(R_t) = S_{0,r} + S_{1,r}\psi_r\left(S_{2,r}f_t\right),$$
$$S_{1,d} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, S_{2,d}' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$
$$S_{0,r} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, S_{1,r} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$
$$S_{2,r}' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\psi_d(x) = \exp{(x)}, \psi_r(x) = \tanh{(x)}.$$

The Jacobians of $\psi_d\left(S_{2,d}f_t\right)$ and $\psi_r\left(S_{2,r}f_t\right)$ are

$$\Psi_{d,t} = D_t, \Psi_{r,t}(x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \varkappa_{p\mu t} & \varkappa_{n\mu t} & \varkappa_{pnt} \end{bmatrix} \begin{bmatrix} 1-\pi_{p\mu t}^2 & 0 & 0 \\ 0 & 1-\pi_{n\mu t}^2 & 0 \\ 0 & 0 & 1-\pi_{pnt}^2 \end{bmatrix},$$

where

$$\varkappa_{p\mu t} = \pi_{n\mu t} - \pi_{p\mu t}\pi_{pnt}\sqrt{\frac{1-\pi_{n\mu t}^2}{1-\pi_{p\mu t}^2}}, \varkappa_{n\mu t} = \pi_{p\mu t} - \pi_{n\mu t}\pi_{pnt}\sqrt{\frac{1-\pi_{p\mu t}^2}{1-\pi_{n\mu t}^2}},$$
$$\varkappa_{pnt} = \sqrt{\left(1-\pi_{p\mu t}^2\right)\left(1-\pi_{n\mu t}^2\right)},$$
$$\dot{D}_t = S_{1,d}\Psi_{d,t}S_{2,d}, \dot{R}_t = S_{1,r}\Psi_{r,t}S_{2,r}.$$

We estimate the score-driven state-space model by maximum likelihood. For the starting values in the optimization, we use the estimates of the coefficients from the baseline state-space model. The estimated covariance parameters from the baseline model (after transformation) are used as the starting f_0 in the dynamic Equation 97. We set the starting values A_0, B_0, and c_0 in Equation 97 as follows: A_0 is a diagonal matrix with 0.5 on the diagonal, B_0 is a diagonal matrix with 0.015 on the diagonal, and $c_0 = (1 - A_0)f_0$, ensuring that since the score has a mean of $0, f_t$ will have its mean centred at f_0.

The score-driven state-space model updates as follows. Every period, we start with a one-step-ahead forecast for the state $\alpha_{t|t-1}$, its variance $P_{t|t-1}$, and f_t. Then we can construct the forecast error v_t and its variance F_t to form the contribution to likelihood for that period as

$$-\frac{1}{2}\left(2\log(2\pi) + \log\left(\det(F_t)\right) + v_t' F_t^{-1} v_t. \qquad (100)\right.$$

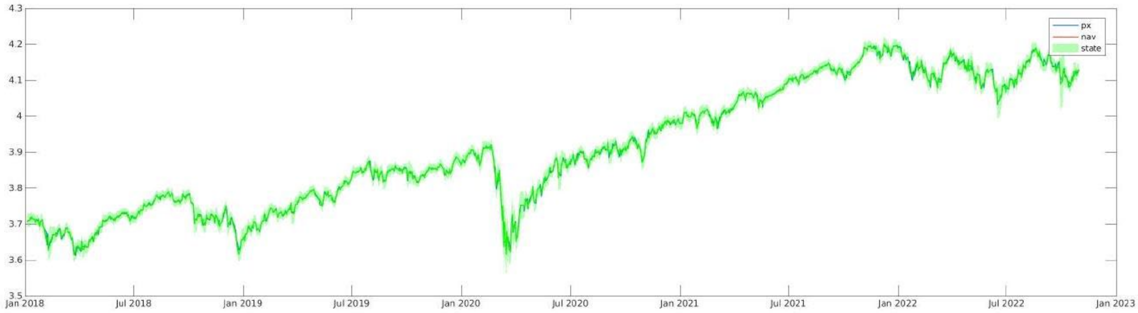
We can then calculate the score and information matrix using Equation 98 and Equation 99[11]. These are used to update f_t for the next period following Equation 97. The next period's f_{t+1} can be transformed to form Q_{t+1} and feed into our updates for the means and variances for the next period $\alpha_{t+1|t}$ and $P_{t+1|t}$. This process continues until it reaches the end of the sample. MATLAB’s Patternsearch algorithm[12] that searches over combinations of the parameters will deliver our estimates for the parameters.

Now we can observe the fitted Price, NAV, and fundamental value chart that are derived using the score-driven model. We can notice from Figure 13 that the standardized residuals are closer to white noise processes and the ACF plots show much less serial correlation under the new estimation method. More generally, comparing Table 4 with Table 3, we find improvements for the heteroskedasticity and serial autocorrelation tests while the normality tests still suffer due to the existence of some outliers. One can also observe that the score model don't solve the heteroskedasticity and serial autocorrelation in the residuals completely. A feature of the model is that when a big shock occurs, the time-varying covariance only gets updated the next period due to the specified dynamics that time-varying parameters follow. This means that there will always be large, standardized residuals for the first period when the shock hits. This explains why even after allowing for time-varying covariance terms, we still occasionally see spikes in the standardized residual plots.

[11] Since the model requires the calculation of the inverse of the information matrix and our time-varying parameter space (6 paramters) is larger than the dimensions of F (2-by-2), we use linear shrinkage towards an identity matrix for the first period and the following period's information matrix would be linearly combined with the information matrix from the previous period. To achieve further stability in the score, we also smooth over the scores by applying a linear combination of the current score with the sum of all previous period's score. This is similar to the idea of momentum in the optimization literature to reduce oscillations and escape local mimima. The weighting used to shrink the information matrix and that used to smooth scores are estimated from the data, along with the other parameters.

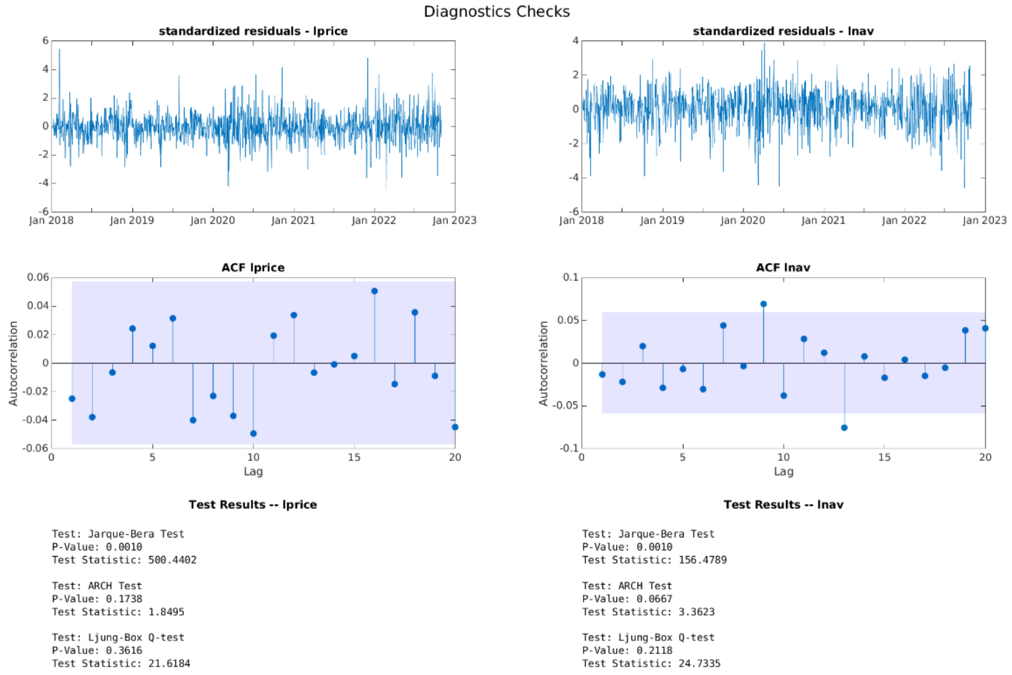
[12] The patternsearch algorithm does direct grid search that doesn't require gradients of the objective function. It's particularly beneficial for nonsmooth, discontinuous and noisy objective functions where traditional gradient-based methods fail. We also experimented with other optimization methods including L-BFGS, SQP, and Chris Sim's Csminwel and found Patternsearch works the best for our model in simulations.

Figure 13: Time-varying Parameters Fitted Model for iShares Core MSCI World UCITS ETF



(a) Price, NAV, Fundamental Value Chart

The above figure demonstrates the ETF price (blue), NAV (red), and the filtered first state variable (bright green) that corresponds to M_t or μ_t. The bands around the green line denote the 95% confidence interval constructed using the state forecast variance for each time period.



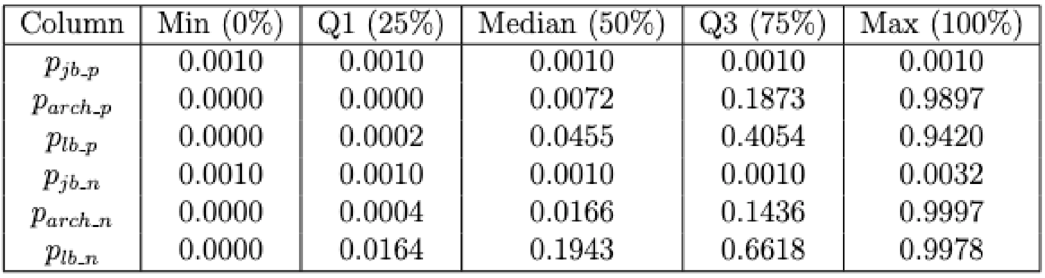
```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 500.4402

Test: ARCH Test
P-Value: 0.1738
Test Statistic: 1.8495

Test: Ljung-Box Q-test
P-Value: 0.3616
Test Statistic: 21.6184
```

```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 156.4789

Test: ARCH Test
P-Value: 0.0667
Test Statistic: 3.3623

Test: Ljung-Box Q-test
P-Value: 0.2118
Test Statistic: 24.7335
```

(b) Diagnostics

The above demonstrates the diagnostic checks for the score-driven time-varying-parameter state-space model. The first row shows the standardized forecast residuals for predicting the two variables in the measurement equation: $log\ (\ price\ )$ and $log\ (NAV)$. The second row shows the ACF plots for the standardized residuals along with Bartlett's confidence interval. The third row presents the test statistics and p-values for the two measurement residuals: normality test, heteroskedasticity test, and autocorrelation test.

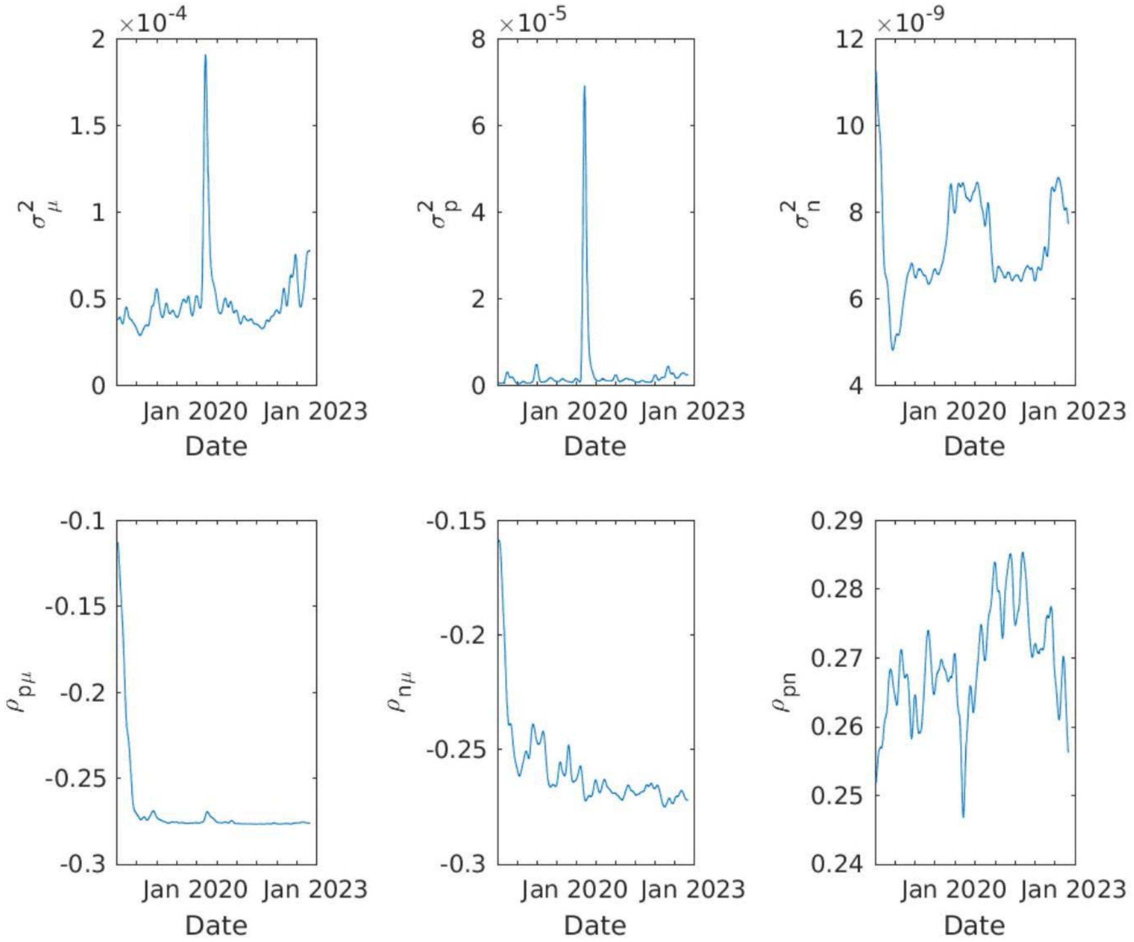
Table 4: Summary Diagnostics for Sample ETFs from the score model



Column	Min (0%)	Q1 (25%)	Median (50%)	Q3 (75%)	Max (100%)
p_{jb_p}	0.0010	0.0010	0.0010	0.0010	0.0010
p_{arch_p}	0.0000	0.0000	0.0072	0.1873	0.9897
p_{lb_p}	0.0000	0.0002	0.0455	0.4054	0.9420
p_{jb_n}	0.0010	0.0010	0.0010	0.0010	0.0032
p_{arch_n}	0.0000	0.0004	0.0166	0.1436	0.9997
p_{lb_n}	0.0000	0.0164	0.1943	0.6618	0.9978

As we did before for the nonparametric estimation, we can now extract the mean time-varying covariance terms from the estimated score model. These are presented in Figure 14. There are a few things worth noting. The significant jump in the error variances also appeared for σ_u^2 and σ_p^2, but the jump is less significant for σ_n^2. This is reasonable as the ETF market is usually more liquid than its underlying and experiences much more volatility and market microstructure noises. The NAV, on the other hand, has been known to be stale due to less trading, particularly for illiquid instruments. This feature is also evident from observing the large jumps and quick overreactions in the price time series, while this is less common in the NAV time series. In terms of the correlations, the score model took a few initial periods to settle down on a relatively stable level of correlation $\rho_{p\mu}$ and $\rho_{n\mu}$ (around -0.25) for the whole sample period, unlike the large variations observed in the nonparametric model. On the other hand, we observe a consistent result of a significant drop in correlation between price and NAV when Covid hit due to hedging.

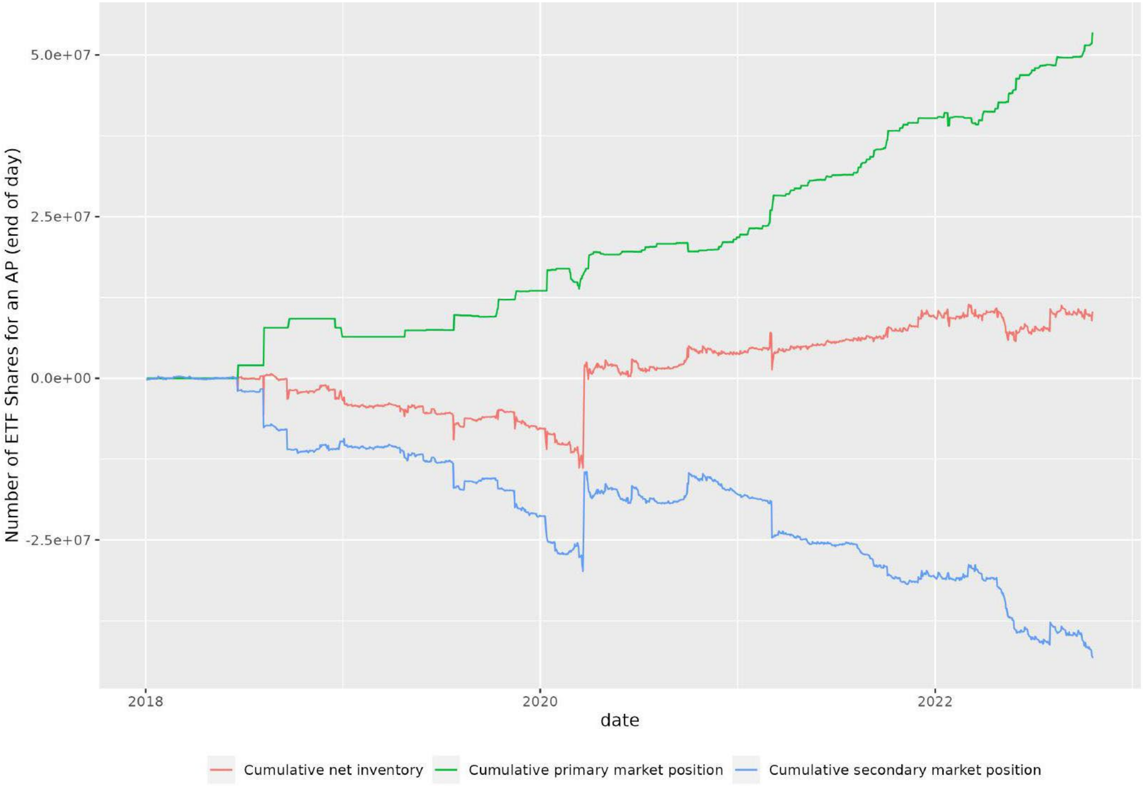
Figure 14: Mean Time Varying Error Covariance (Score Model). across sample ETFs



5.2 Time-Varying Optimal Inventory

Our previously derived model suggests an ARIMA dynamic (Equation 82) for the AP's inventory level. However, the observed time series of the AP's inventory suffer from the presence of outliers. One example is a plot of the secondary, primary cumulative net positions, as well as the cumulative inventory of a representative AP in the representative ETF: Figure 15. Notice the occasional jumps in the inventory time series that look discordant from the most observations.

Figure 15: Representative AP's Inventory



There are two potential reasons behind the outliers. Firstly, the dynamic in Equation 82 is derived assuming a static I^d. In reality, I^d could change once in a while depending on the changing covariance between the ETF's fundamental value and outside income, as described in Section 3.2, over the 5-year sample period. The ETF inventory that we observe is mostly only a part of a larger, unobserved portfolio that the AP is holding. These unobserved changes to the optimal inventory could take different forms: a one-off additive shock (AO), a permanent shift (LS), a temporary change (TC) that gradually diminishes, a seasonal shift (SLS) (due to rebalancing or tax reasons), or a shock to the error term (IO) of the inventory dynamic (specifically referring to the predicted noise trader's demand x in our model), whose effects propagate through future inventories according to the specified model. Secondly, while our data should cover the majority of the secondary market for these ETFs, which are primarily traded on the LSE, there may be some unobserved trades occurring on other exchanges where none of the involved parties have reporting obligations to the FCA. These missing trades could manifest as artificial deviations in the observed inventory levels.

It would be easy to control for these kinds of shifts in optimal inventory if we knew the timing and patterns of the shifts; then we would separate these shifts (due to outside/unobserved factors) from the usual daily inventory dynamics of the AP as a market maker. Since the above information isn't available to us, we can statistically detect any deviation from the expected inventory dynamics by examining the correlation structures in the residuals and categorize them into different types of "outliers"[13]. For

[13] As we follow Madhavan and Smidt (1993) in selecting around 5 outliers per year (adding up to 20-30 outliers per AP's inventory time series), it may not be justifiable to refer to these effects as "outliers" given the number selected. Instead, it is more accurate to view them as interventions in the AP's inventory management, resulting in deviations from the theoretical $ARIMA$ process.

this purpose, we follow the outlier detection technique introduced in Chen and Liu (1993) that can detect different types of outliers including additive (AO), level shifts (LS), temporary changes (TC), seasonal level shifts (SLS) and innovational (IO).

Assuming the AP's outlier-free inventory (e.g., when the optimal level of inventory is zero) is denoted as $\tilde{I}_t$ and follows the following process

$$\frac{\phi(L)\alpha(L)}{\theta(L)}\tilde{I}_t = \Psi_I(L)\tilde{I}_t = \epsilon_t, \tag{101}$$

where L is the lag operator, and $\phi(L), \theta(L)$ and $\alpha(L)$ are the lag polynomials representing the AR, MA, and integrated part of the general ARIMA process. The observed inventory I_t, subject to different kinds of interventions, is then defined as

$$I_t = \tilde{I}_t + sDe_t(t^*), \tag{102}$$

where $\tilde{I}_t$ follows the $AR(1)$ process defined above. $e_t(t^*)$ is an indicator function for the occurrence of the intervention, where $e_t(t^*) = 1$ if $t = t^*$ and $e_t(t^*) = 0$ otherwise. s denotes the magnitude of the intervention, and D denotes the dynamics of the intervention depending on its type. The types mentioned in the previous paragraph correspond to the following cases

$$AO: D = 1, \tag{103}$$

$$LS: D = \frac{1}{1-L}, \tag{104}$$

$$TC: D = \frac{1}{1-\delta L}, \tag{105}$$

$$SLS: D = \frac{1}{1-B^k}, \tag{106}$$

$$IO: D = \frac{1}{\Psi_I(L)}. \tag{107}$$

To test if an observation is an outlier following Tsay (1986), we first note that Equation 102 can be rewritten as

$$\varepsilon_t = \Psi_I(L)I_t = \epsilon_t + s\Psi_I(L)De_t(t^*). \tag{108}$$

In the above equation, both $\Psi_I(L)I_t$ and $\Psi_I(L)De_t(t^*)$ are known values, given the model parameters and assuming the current observation is an outlier of a certain type. Therefore, we can estimate the magnitude of the effect $\hat{s}$ using least squares. For example, $\hat{s}_{AO,t} = \varepsilon_t$, and $\hat{s}_{IO,t} = \frac{1}{1+\phi_I^2}(1-\phi_I L)\varepsilon_t$. We can also derive the variances for these $\hat{s}$ given an estimate for $\sigma^2_{\varepsilon_t}$. With estimators for the model parameters and error variances, we can then form test statistics for each type of outlier for the current observation. The implementation for the detection is as follows.

- Step 1: Fit the time series using auto arima (which select the best model that fit the data according to information criteria) and obtain the initial model parameters and residuals.

- Step 2: For each time period, we calculate the 5 test statistics (using the residuals) for testing whether or not the observation for the period is an outlier of each type, as discussed in the previous paragraph. Among the $5T$ test statistics calculated, we find the biggest one and if this test statistics' magnitude is bigger than a criterion, then it is classified as a potential outlier. The effect of this outlier is removed from the residuals and the current step is repeated until no more outliers are detected under the current model parameters.

- Step 3: Controlling for all identified outliers and re-fit an auto arima model as in Step 1. Repeat Step 2 with the new residual series. Iterate between step 1 (model estimation) and Step 2 (outlier detection) until no additional outliers can be identified under a given model.

- Step 4: Once we have our pool of potential outliers, we proceed to refine them. We start with the biggest model, and jointly estimate the effects of all m identified outliers from the previous two steps using

  $$\varepsilon_t = \epsilon_t + \sum_{j=1}^{m} s_j \Psi_I(L) D_j e_t(t_j), \tag{109}$$

  as each period may be affected by the j^{th} outliers occurred in period t_j. We can then construct the t statistics for each of the outlier effects s_j. If the minimum of these t statistics is below the criterion value, we then eliminate the outlier from the pool and re-run the regression in Equation 109. This step iterates until all the remining outliers are significant.

An illustration of the outlier detection results is shown in Figure 16. This figure shows a time series of one AP's inventory in a ETF. The top chart shows the original time series in grey and the corrected time series in blue while the bottom chart shows the outlier process detected. Table 5 lists out all the types of outliers detected in the sample, along with the time of occurrence, magnitude, and t-statistics. The biggest jumps in this AP's inventory are the two AOs that occurred near the end of the sample that quickly corrected itself. Also noticeable are the IO and TC that take several periods to die out. The inventory time series is much more reasonable to be used in later empirical analysis and the outlier process will be taken as the interventions to the optimal inventory level (either due to changing covariances with outside income or other unobserved trading activities our dataset is not capturing).

Figure 16: Outlier detection illustration

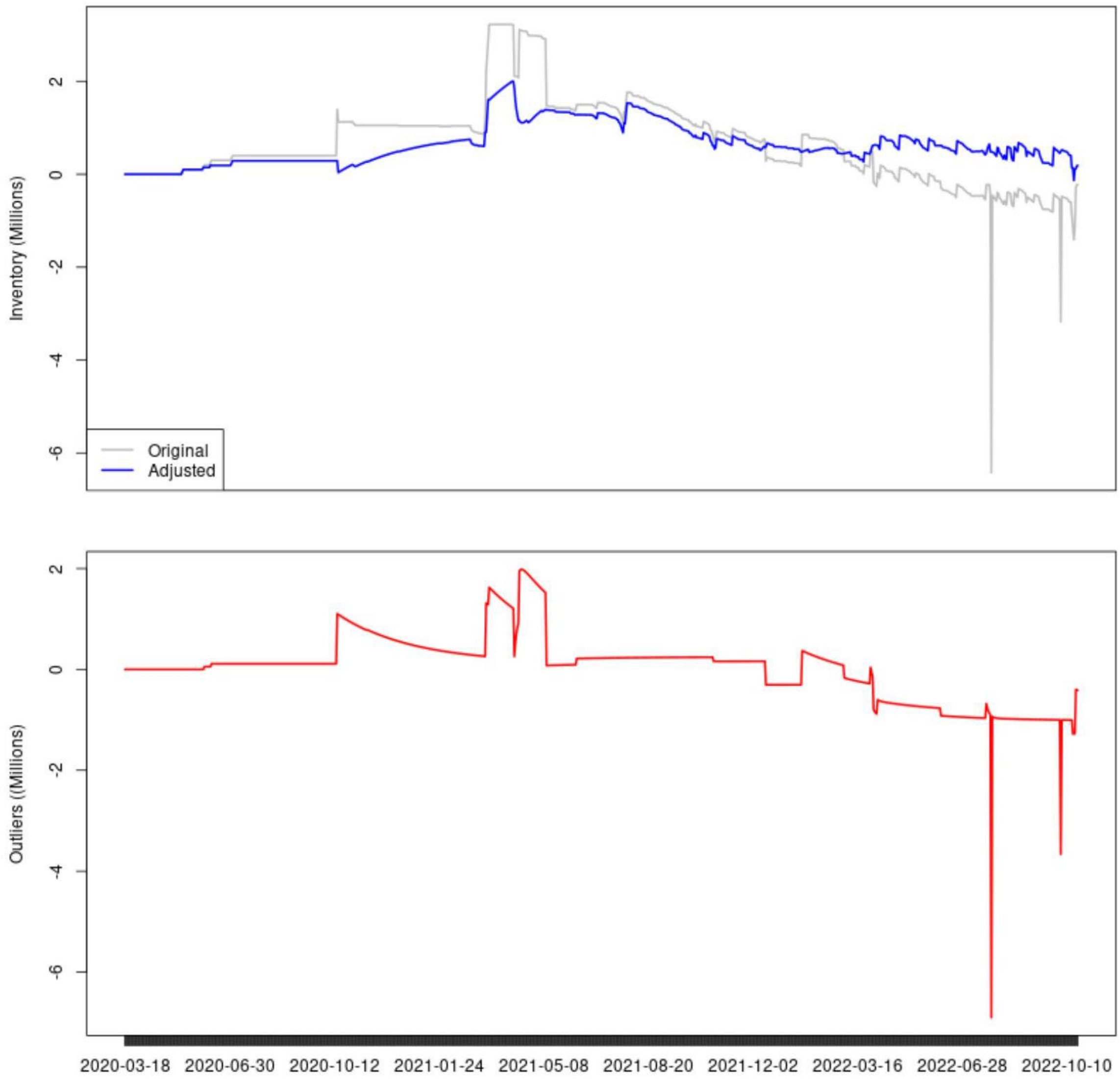
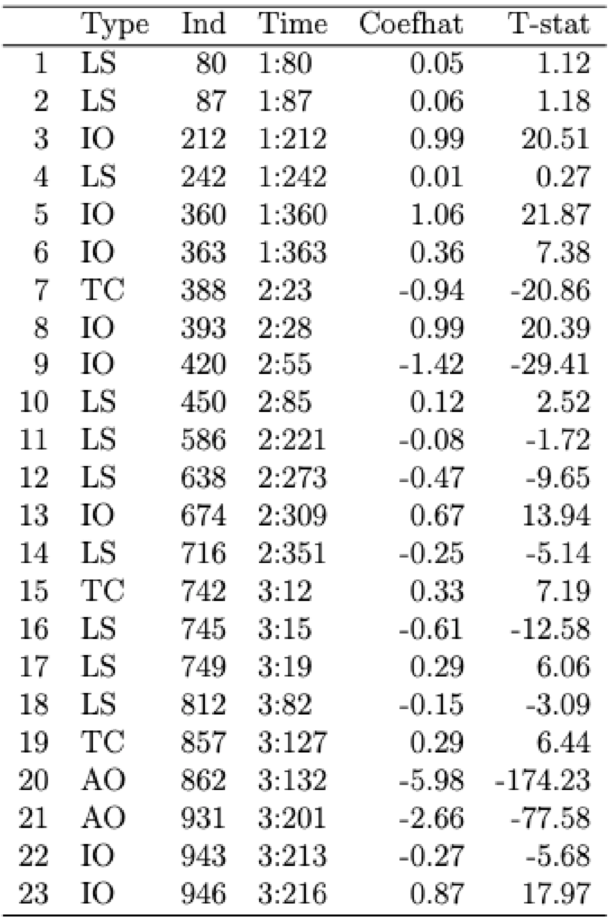


Table 5: Outlier detection illustration



	Type	Ind	Time	Coefhat	T-stat
1	LS	80	1:80	0.05	1.12
2	LS	87	1:87	0.06	1.18
3	IO	212	1:212	0.99	20.51
4	LS	242	1:242	0.01	0.27
5	IO	360	1:360	1.06	21.87
6	IO	363	1:363	0.36	7.38
7	TC	388	2:23	-0.94	-20.86
8	IO	393	2:28	0.99	20.39
9	IO	420	2:55	-1.42	-29.41
10	LS	450	2:85	0.12	2.52
11	LS	586	2:221	-0.08	-1.72
12	LS	638	2:273	-0.47	-9.65
13	IO	674	2:309	0.67	13.94
14	LS	716	2:351	-0.25	-5.14
15	TC	742	3:12	0.33	7.19
16	LS	745	3:15	-0.61	-12.58
17	LS	749	3:19	0.29	6.06
18	LS	812	3:82	-0.15	-3.09
19	TC	857	3:127	0.29	6.44
20	AO	862	3:132	-5.98	-174.23
21	AO	931	3:201	-2.66	-77.58
22	IO	943	3:213	-0.27	-5.68
23	IO	946	3:216	0.87	17.97

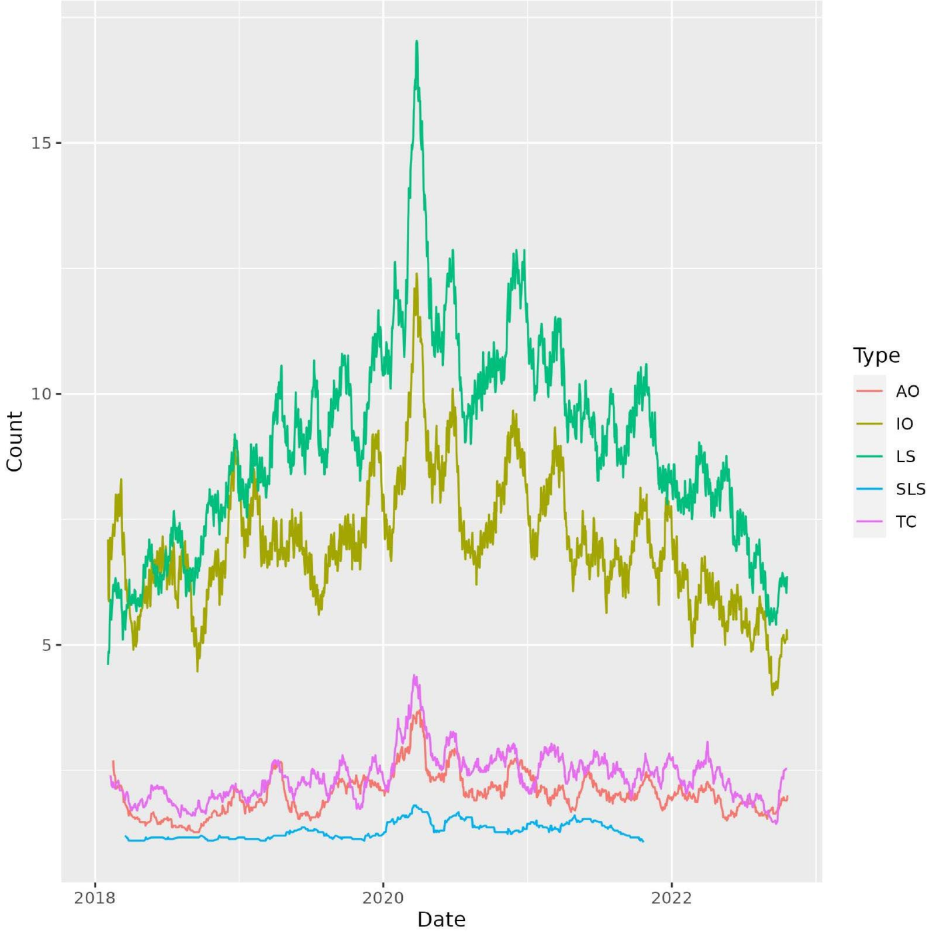
Outlier Detection Results

After running the outlier detection for each ETF-AP pair in our sample, we can construct Figure 17, which represents the 30-day moving average of the total number of different types of outliers detected for each day across all ETF-AP pairs.

We observe that all outlier types grow significantly during the Covid period, particularly for LS (Level Shift) and IO (Innovational Outliers). These two types of outliers indicate fundamental changes in an AP's inventory level, either a permanent shift (LS) or a shock to the error term (IO) that will gradually dissipate if the outlier-free model is persistent.

This observation is crucial because it suggests that a short-lived Covid market turmoil may have more long-lasting impacts on the AP's inventory management mechanisms. The detection of LS and IO during this period highlights how APs may have undergone structural adjustments in response to the increased uncertainty and market volatility, leading to persistent changes in their behaviour even after the immediate crisis passed.

Figure 17: Outlier types detected during the sample period



Note: This figure represents the 30-day moving average of the number of different types of outliers detected on each day.

5.3 Daily Order Imbalance

From the price and inventory dynamics derived in Equations 66 and 61, we recognise that an essential factor in the AP's price and inventory decisions is the excess market demand. In Equation 66, the noise trader demand can be interpreted in two ways. One interpretation is the exogenous order imbalance that acts as a shock that the AP, as a market maker, must accommodate. Alternatively, it can be seen as the unexpected component of the market's excess demand.

The AP forms expectations about the fundamental value each period and learns about the informed trader's optimal demand function over time. Therefore, the expected informed trader demand also represents their expected market excess demand at a given price. The difference between the observed realized market excess demand and the expected informed trader demand is the "unexplained" portion, which can only be attributed to noise traders. In Equation 66, the AP updates its belief on the fundamental value of the ETF from observing this "unexpected" component of market excess demand.

This second interpretation motivates us to estimate the noise traders' demand by measuring the unexpected part of the market's excess demand. To first calculate the market excess demand, we sign each observed transaction as buyer- or seller-initiated based on the tick rule proposed by Harris (1989). If the transaction price is higher than the previous price, we assign it as "buyer-initiated". If it is lower than the previous price, we assign it as "seller-initiated". If it equals the previous price, the transaction inherits the direction from the prior transaction.

Given that our secondary market transaction data spans multiple exchanges, each with varying liquidity and latency characteristics, we restrict the comparison to the previous price on the same exchange.

After signing each transaction, we aggregate up the buyer-initiated and seller-initiated trades' volume and take a difference, resulting in a measure of the daily order imbalance $ordimb_t$. Note that we exclude off-exchange trades and large block trades (top and bottom 5% of the intraday transaction volume distribution), as the former cannot have trade direction assigned, and the latter are typically privately negotiated in the upstairs market, not reacting to the AP's pricing in the same way as other orders on public exchanges.

We then use the observed daily order imbalance to estimate the predicted daily noise traders' demand (or, alternatively, the "surprise" component of daily observed excess demand) by running the following regression. This formulation and lag order selections follow the methodology of Madhavan and Smidt (1993) and Hasbrouck (1991) in their studies of NYSE stocks and specialist trades. The residuals from the regression below represent the unexpected part of the order imbalance, which we use for later empirical analysis

$$ordimb_t = \gamma_0 + \sum_{i=1}^{3} \gamma_i ordimb_{t-i} + \sum_{j=1}^{3} \delta_j \left(p_{t-j} - p_{t-j-1}\right) + x_t. \tag{110}$$

5.4 Heterogeneous APs

The APs in our sample include a variety of firms, such as high-frequency traders and banks. These firms likely have different inventory tolerances/costs, leading to varying degrees of deviation from their optimal inventory levels. It is also reasonable to believe they differ in terms of available outside income sources, which could lead to distinct patterns in the time-varying optimal inventory levels[14]. To explore the heterogeneity of the effect of inventory management on ETF mispricing for different types of APs, we aim to classify the APs into several distinct groups.

Additionally, our primary market data includes APs' LEIs at the subsidiary level, either geographical (e.g., Flow Trader London Ltd) or functional (e.g., J.P. Morgan Asset Management). This implies that classifying the APs based on their parent firm's category may not be reasonable, as different desks within the same firm could have separate accounts and different objectives. Therefore, we believe it is more reliable to use transaction data to help us identify AP groupings.

[14] Past literature has attempted to distinguish globally systemically important banks (G-SIBs) from other APs (see Gorbatikov and Sikorskaya (2022)), as they face higher regulatory costs that result in greater balance sheet constraints for holding inventory.

We expect to categorize the APs in each ETF market into 3 groups based on their trading behaviour: Market Makers, High-Frequency Traders, and Investment Banks.

Market Makers and High-Frequency Traders for each ETF are intraday intermediaries who engage in frequent buying and selling throughout the day to meet liquidity needs from buyers and sellers (See Grossman and Miller (1988); Glosten and Milgrom (1985)). They exhibit a fast speed of mean reversion in inventory (See Amihud and Mendelson (1986); Ho and Stoll (1983); Madhavan and Smidt (1993)), while maintaining a low end-of-day net position (Kirilenko et al. (2017)). High-Frequency Traders are the most active subgroup of intraday intermediaries, with particularly short between-trade durations and high trading frequency. Investment Banks, when acting as APs in the ETF market, often focus on servicing large institutional clients, and don't engage in the ETF market with the same frequency as the prior two groups but may hold large positions over time.

Based on the features of these different groups of traders, we collect the following variables summarizing their daily trading behaviour: volume (number of ETF shares traded in the secondary market), trades per hour (number of daily trades divided by the daily trading hours), median duration (median duration between two consecutive trades), cumulative end-of-day position (adjusted for outliers), cumulative outlier process, and cumulative end-of-day primary market position. To achieve a data-driven classification that evolves smoothly through time, we adopt the Smoothplaid biclustering algorithm proposed in Mankad et al. (2013).

For each trading day, we construct a data matrix where each row represents an Authorised Participant (AP) and each column corresponds to one of six trading features. Using biclustering, we identify clusters of traders and features that exhibit similar patterns. Unlike traditional clustering, which groups data based on overall similarity, biclustering is flexible enough to capture traders who behave similarly with respect to only a subset of features. After identifying the clusters, we infer the values of each trader-feature pair by fitting the data to a cluster-based fixed-effects model, generating denoised estimates that amplify meaningful patterns and reduce noise. This approach enhances the classification of traders and their trading behaviours. We apply the SmoothPlaid algorithm developed in Mankad et al. (2013), which improves upon traditional biclustering by introducing smoothness penalties across time, ensuring that clusters evolve gradually rather than abruptly.
Specifically, in our model, each bicluster consists of a common effect, a trader-specific effect, and a feature-specific effect. For a daily data matrix X, where each column is standardized to have a mean of zero and a standard deviation of one, the value of an element X_{ij} is represented as

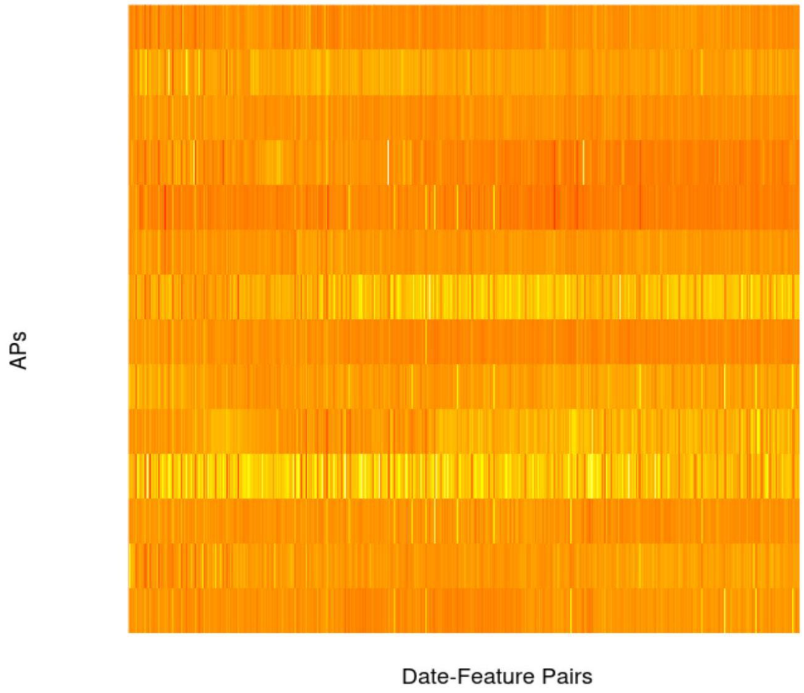
$$X_{ij} = \mu_0 + \sum_{k=1}^{K} \theta_{ijk} r_{ik} c_{jk},$$

where $i = 1, \dots, n$ indexes traders, $j = 1, \dots, p$ indexes feature, K is the number of layers, and r_{ik} and c_{jk} are indicators of whether the ij^{th} element belongs to cluster k. The base layer effect, μ_0, is the global mean of the data matrix, and θ_{ijk} is the element-specific effect in cluster k, which is characterized as

$$\theta_{ijk} = \mu_k + \alpha_{ik} + \beta_{jk},$$

where μ_k, α_{ik}, and β_{jk} represent the common effect, trader-specific effect, and feature-specific effect for layer k. This structure assumes traders within the same cluster exhibit similar strategies, with individual differences across features. One example of the fitted X matrix for APs in the iShares Core MSCI World UCITS ETF market is shown in Figure 18.

Figure 18: Biclustering matrix for iShares Core MSCI World UCITS ETF

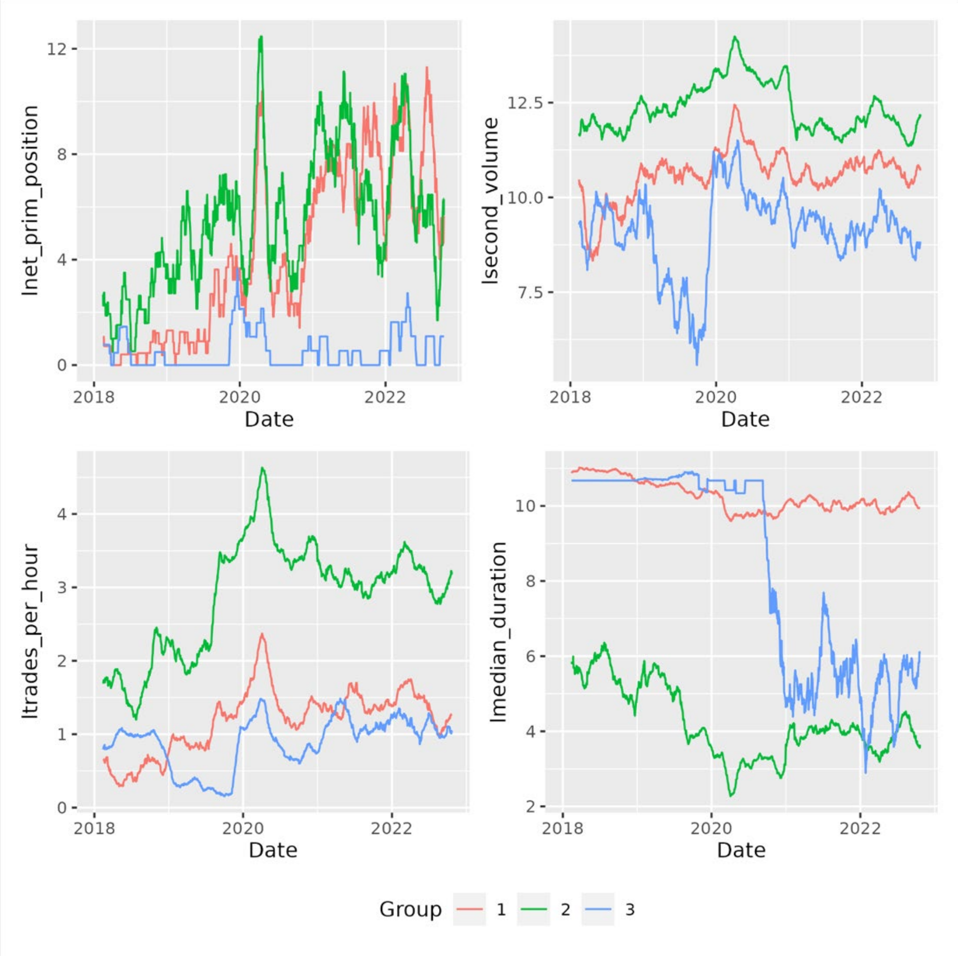


Biclusters are estimated sequentially, with the K^{th} layer fitted to the residuals of the previous layers

$$\hat{Z}_{ij} = X_{ij} - \hat{\mu}_0 - \sum_{k=1}^{K} \hat{\theta}_{ijk}\hat{r}_{ik}\hat{c}_{jk}.$$

We then applied hierarchical clustering to group APs based on their denoised fitted values, identifying distinct clusters of APs with similar trading behaviours. Figure 19 displays the mean values for each of the four behavioural features across the three identified groups for a representative ETF. Note that we obtained the grouping using the biclustering techniques shown above but the mean value shown in Figure 19 is calculated using original data. Group 2 appears to consist of high-frequency traders as they are most active in both the secondary and primary market, have the most number of trades per hour and shortest inter-trade duration. Group 1 likely represents general market makers as they have a consistent presence in both markets while conducting fewer trades per hour and have longer duration compared with a high frequency trader. Group 3 likely includes investment banks who occasionally engage in the ETF market based on client needs. It is interesting to note that during the Covid turmoil period and the volatile period near the beginning of 2022, the cumulative primary market positions of Group 1 and Group 2 APs went through volatile jumps, which correspond to high frequency traders and market makers satisfying volatile market demand. These two groups also share similar time series patterns in all the features except for median trading duration.

Figure 19: Heterogeneous APs for iShares Core MSCI World UCITS ETF



In the empirical section, we experimented with samples consisting of different groups of APs, and we find our model implications are mostly demonstrated by the combinations of the first two types of AP - market makers and high-frequency traders. Therefore, the results shown in Section 6 will be based on the aggregate behaviour of these two groups of APs for each ETF.

6 Empirical Findings

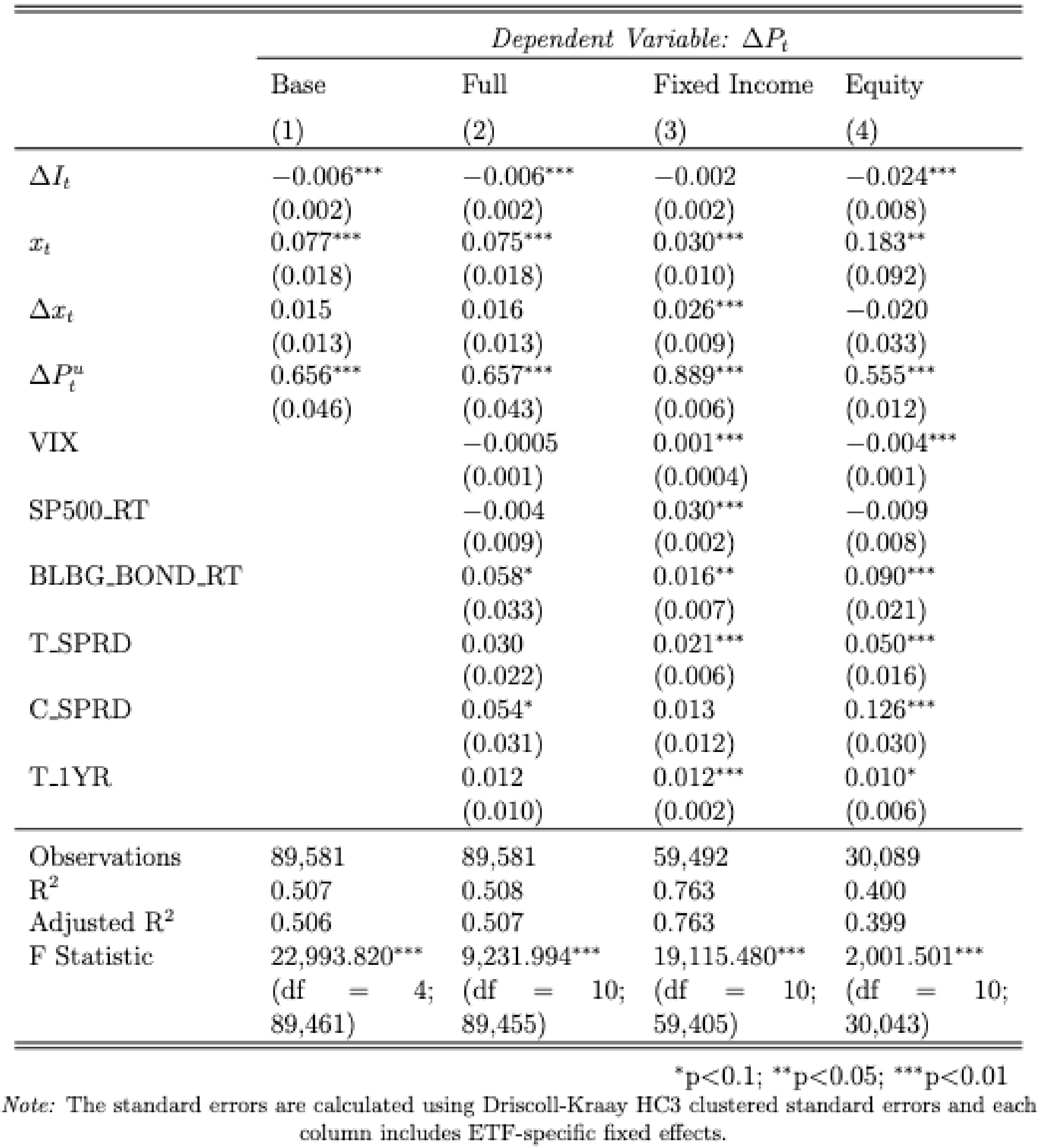
In this section, we demonstrate empirical evidence supporting the dynamics derived in Section 3. Specifically, we utilise the observed ETF price, NAV, AP's inventory, as well as the extracted ETF fundamental values from Subsection 5.1, AP's optimal level of inventory from Subsection 5.2, and the market's unexpected order imbalance from Subsection 5.3, in our regression analysis. For each dynamic derived in the model, we assume that the real data contains some measurement noise in the variables, and we account for this by adding an error term to form our regressions. All the regressions are estimated using panel data models with individual fixed effects and we used clustered standard errors Hoechle (2007) to take into account the cross-sectional dependence of sample ETFs and temporal dependence within each ETF.

6.1 Price Dynamics

$$\Delta P_t = (1-\alpha_3)\frac{\Omega}{\delta(1-\Omega)}x_t + \alpha_1 \Delta I_t + \alpha_2 \Delta x_t + \alpha_3 \Delta P_t^u + \epsilon_{pt}, \tag{111}$$

where in the model $\alpha_1 < 0, \alpha_2 > 0, 0 < \alpha_3 < 1$.
We estimate our price dynamic using a panel regression across all ETFs in our sample period. The results are presented in Table 6. We observe that even after controlling for common macroeconomic variables, the coefficients α_1, α_2, and α_3 are mostly significant and display the expected signs. The exceptions are that changes in inventory do not significantly affect the prices of Fixed Income ETFs, and changes in market order imbalance do not significantly affect Equity ETFs. This indicates that Fixed Income ETF prices are more responsive to market order imbalance, whereas Equity ETFs' prices respond more to APs' inventory management. Changes in NAV have a positive effect on ETF price changes, and the combined effects of unexpected order imbalance and NAV changes sum to approximately 1, as the model predicts.

Table 6: Price Dynamics



	Dependent Variable: ΔP_t			
	Base	Full	Fixed Income	Equity
	(1)	(2)	(3)	(4)
ΔI_t	-0.006^{***}	-0.006^{***}	-0.002	-0.024^{***}
	(0.002)	(0.002)	(0.002)	(0.008)
x_t	0.077^{***}	0.075^{***}	0.030^{***}	0.183^{**}
	(0.018)	(0.018)	(0.010)	(0.092)
Δx_t	0.015	0.016	0.026^{***}	-0.020
	(0.013)	(0.013)	(0.009)	(0.033)
ΔP^u_t	0.656^{***}	0.657^{***}	0.889^{***}	0.555^{***}
	(0.046)	(0.043)	(0.006)	(0.012)
VIX		-0.0005	0.001^{***}	-0.004^{***}
		(0.001)	(0.0004)	(0.001)
SP500_RT		-0.004	0.030^{***}	-0.009
		(0.009)	(0.002)	(0.008)
BLBG_BOND_RT		0.058^{*}	0.016^{**}	0.090^{***}
		(0.033)	(0.007)	(0.021)
T_SPRD		0.030	0.021^{***}	0.050^{***}
		(0.022)	(0.006)	(0.016)
C_SPRD		0.054^{*}	0.013	0.126^{***}
		(0.031)	(0.012)	(0.030)
T_1YR		0.012	0.012^{***}	0.010^{*}
		(0.010)	(0.002)	(0.006)
Observations	89,581	89,581	59,492	30,089
R^2	0.507	0.508	0.763	0.400
Adjusted R^2	0.506	0.507	0.763	0.399
F Statistic	$22{,}993.820^{***}$ (df = 4; 89,461)	$9{,}231.994^{***}$ (df = 10; 89,455)	$19{,}115.480^{***}$ (df = 10; 59,405)	$2{,}001.501^{***}$ (df = 10; 30,043)

*p<0.1; **p<0.05; ***p<0.01

Note: The standard errors are calculated using Driscoll-Kraay HC3 clustered standard errors and each column includes ETF-specific fixed effects.

6.2 NAV Dynamics

$$\Delta P^u_t = -2\theta_1(I_{t-1} - I^d_{t-1}) + \theta_2(P^u_{t-1} - \mu_{t-1}) + \theta_1 x_t + \epsilon_{nt}, \qquad (112)$$

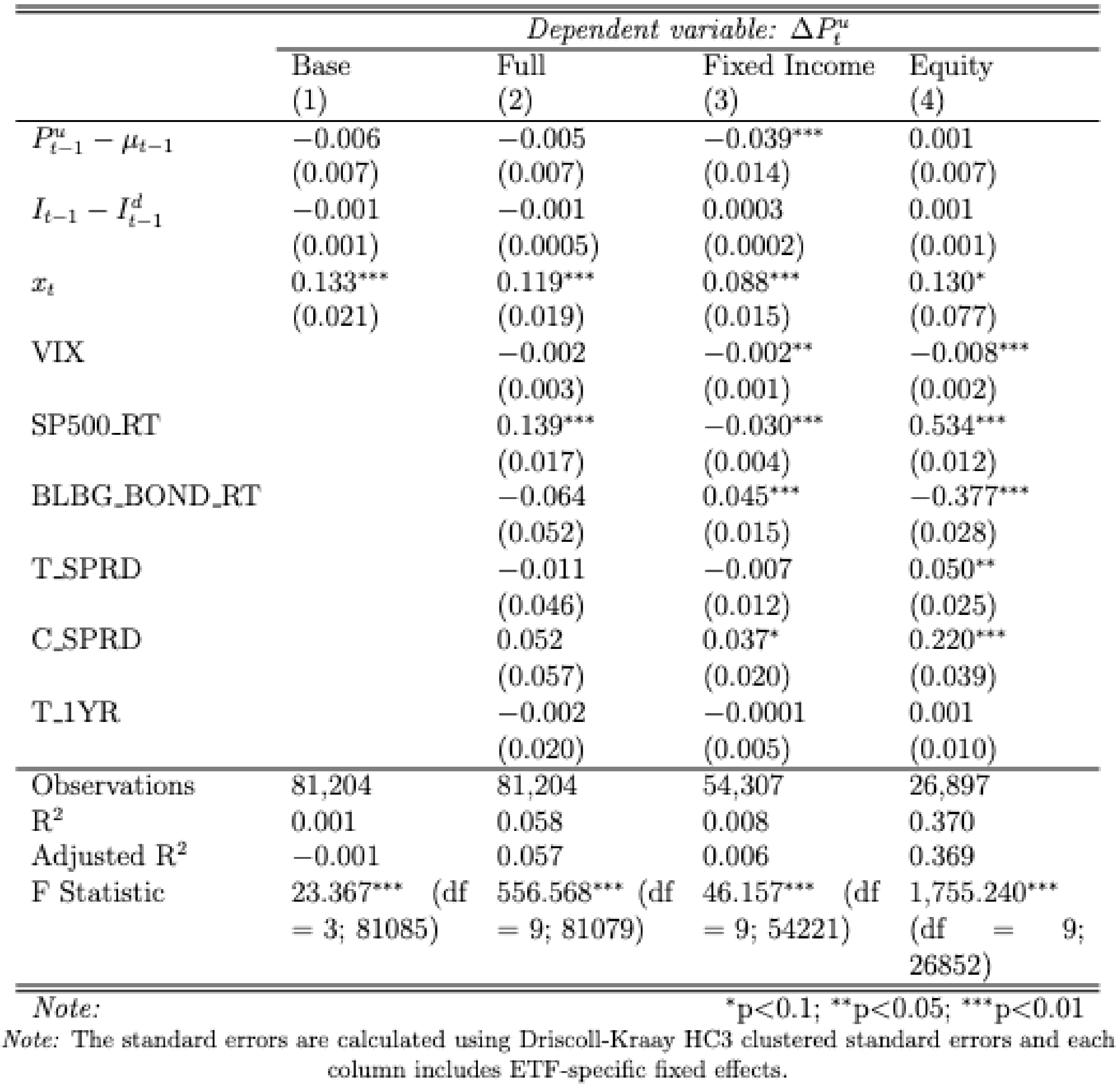
where, according to the model, $\theta_1 > 0$ and $\theta_2 < 0$.
The panel regression results for the NAV dynamic are reported in Table 7. The unexpected order imbalance x_t has a significantly positive effect on changes in NAV, which demonstrates the transmission mechanism of shocks between the ETF market and its underlying assets through APs' arbitrage activities. This effect is especially strong for

Fixed Income ETFs, whose opacity and illiquidity make the ETF market an important source for price discovery.

The changes in Fixed Income ETFs' NAVs also respond negatively to the deviation from their fundamental value, indicating a gradual absorption of information that leads to the NAV converging toward the fundamental value. This correction effect is not observed in Equity ETFs, likely due to their higher price efficiency, where adjustments often occur intraday and are therefore not captured in end-of-day NAVs.

The expected impact of APs' inventory situations on changes in NAVs is generally not observed. In general, the model explains little variation in NAV changes. This outcome is not unexpected, given that most ETFs in our sample consist of baskets of hundreds or thousands of underlying assets. As such, the indirect relationship between APs' inventory management and NAV changes may be too diffuse to be easily captured by the model.

Table 7: NAV Dynamics



	Dependent variable: ΔP^u_t			
	Base	Full	Fixed Income	Equity
	(1)	(2)	(3)	(4)
$P^u_{t-1} - \mu_{t-1}$	-0.006	-0.005	-0.039^{***}	0.001
	(0.007)	(0.007)	(0.014)	(0.007)
$I_{t-1} - I^d_{t-1}$	-0.001	-0.001	0.0003	0.001
	(0.001)	(0.0005)	(0.0002)	(0.001)
x_t	0.133^{***}	0.119^{***}	0.088^{***}	0.130^{*}
	(0.021)	(0.019)	(0.015)	(0.077)
VIX		-0.002	-0.002^{**}	-0.008^{***}
		(0.003)	(0.001)	(0.002)
SP500_RT		0.139^{***}	-0.030^{***}	0.534^{***}
		(0.017)	(0.004)	(0.012)
BLBG_BOND_RT		-0.064	0.045^{***}	-0.377^{***}
		(0.052)	(0.015)	(0.028)
T_SPRD		-0.011	-0.007	0.050^{**}
		(0.046)	(0.012)	(0.025)
C_SPRD		0.052	0.037^{*}	0.220^{***}
		(0.057)	(0.020)	(0.039)
T_1YR		-0.002	-0.0001	0.001
		(0.020)	(0.005)	(0.010)
Observations	81,204	81,204	54,307	26,897
R^2	0.001	0.058	0.008	0.370
Adjusted R^2	-0.001	0.057	0.006	0.369
F Statistic	23.367^{***} (df = 3; 81085)	556.568^{***} (df = 9; 81079)	46.157^{***} (df = 9; 54221)	$1,755.240^{***}$ (df = 9; 26852)
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: The standard errors are calculated using Driscoll-Kraay HC3 clustered standard errors and each column includes ETF-specific fixed effects.

6.3 Mispricing Dynamics

$$\pi_t = \tau_1 \pi_{t-1} + \tau_2 (I_t - I_t^d) + \tau_3 (I_{t-1} - I_{t-1}^d) + \tau_4 x_t + \tau_5 x_{t-1} + \tau_6 \frac{\Omega}{1-\Omega} x_t + \epsilon_{\pi,t}, \tag{113}$$

where in the model $\tau_1 > 0, \tau_3 > 0, \tau_4 > 0, \tau_5 < 0, \tau_6 > 0$, and the sign of τ_2 is indeterminate, depending on the counteracting effects from the model for ETF price $(\alpha_1 < 0)$ and NAV $(2\theta_1 > 0)$.

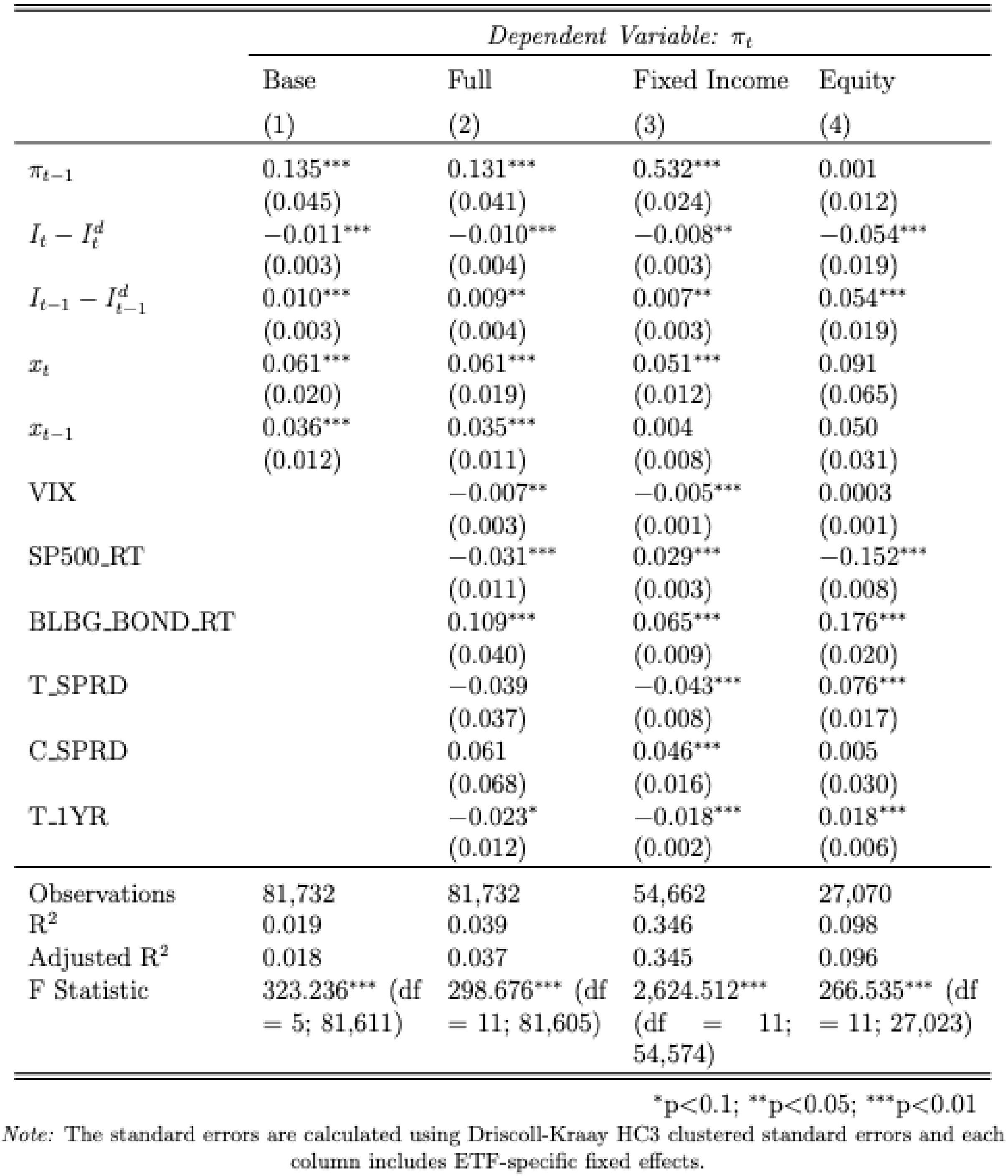
The panel regression results[15] for the mispricing dynamic are presented in Table 8. Lagged mispricing has a strong and significant effect on current mispricing for Fixed Income ETFs, but this effect is insignificant for Equity ETFs. This can likely be attributed to the fact that mispricing corrections in Equity ETFs tend to occur intraday and are not captured by daily data.

Both the current and lagged deviations of inventory from optimal levels significantly impact current mispricing. The lagged inventory effect (τ_3) is positive, as expected from the model, reflecting its negative effect on NAV. The estimated impact of current inventory on mispricing is negative, implying that when APs' inventory exceeds the optimal level, its effect on lowering ETF prices outweighs the effect of decreasing the NAV. This finding is consistent with the limited explanatory power we observed in the model for NAV dynamics.

The current unexpected order imbalance has a significantly positive effect on mispricing for Fixed Income ETFs, but no significant effect is observed for Equity ETFs. This may be related to how closely linked the Equity ETF market is with its underlying assets compared to Fixed Income ETFs. If the unexpected order imbalance is highly correlated in both markets (as is the case for most Equity ETFs), then both the ETF price and NAV may move in the same direction, leading to insignificant effects on mispricing. The lagged unexpected order imbalance does not show any significant effects on mispricing in either subsample, once we control for lagged mispricing and the current inventory situation.

[15] We estimate our dynamic panel data model using standard fixed effects estimation, as we are in a large T, small N context with N=128 and T>1000. The Nickell bias should thus be negligible.

Table 8: Mispricing Dynamics



	Dependent Variable: π_t			
	Base	Full	Fixed Income	Equity
	(1)	(2)	(3)	(4)
π_{t-1}	0.135***	0.131***	0.532***	0.001
	(0.045)	(0.041)	(0.024)	(0.012)
$I_t - I_t^d$	−0.011***	−0.010***	−0.008**	−0.054***
	(0.003)	(0.004)	(0.003)	(0.019)
$I_{t-1} - I_{t-1}^d$	0.010***	0.009**	0.007**	0.054***
	(0.003)	(0.004)	(0.003)	(0.019)
x_t	0.061***	0.061***	0.051***	0.091
	(0.020)	(0.019)	(0.012)	(0.065)
x_{t-1}	0.036***	0.035***	0.004	0.050
	(0.012)	(0.011)	(0.008)	(0.031)
VIX		−0.007**	−0.005***	0.0003
		(0.003)	(0.001)	(0.001)
SP500_RT		−0.031***	0.029***	−0.152***
		(0.011)	(0.003)	(0.008)
BLBG_BOND_RT		0.109***	0.065***	0.176***
		(0.040)	(0.009)	(0.020)
T_SPRD		−0.039	−0.043***	0.076***
		(0.037)	(0.008)	(0.017)
C_SPRD		0.061	0.046***	0.005
		(0.068)	(0.016)	(0.030)
T_1YR		−0.023*	−0.018***	0.018***
		(0.012)	(0.002)	(0.006)
Observations	81,732	81,732	54,662	27,070
R²	0.019	0.039	0.346	0.098
Adjusted R²	0.018	0.037	0.345	0.096
F Statistic	323.236*** (df = 5; 81,611)	298.676*** (df = 11; 81,605)	2,624.512*** (df = 11; 54,574)	266.535*** (df = 11; 27,023)

*p<0.1; **p<0.05; ***p<0.01

Note: The standard errors are calculated using Driscoll-Kraay HC3 clustered standard errors and each column includes ETF-specific fixed effects.

6.4 Inventory Dynamics

$$\Delta I_t = -\iota_1(I_{t-1} - I_{t-1}^d) - \iota_2 x_t - \iota_3(P_{t-1}^u - \mu_{t-1}) + \epsilon_{I,t}, \tag{114}$$

where in the model $\iota_1 > 0, \iota_2 > 0$, and $\iota_3 > 0$.

The panel regression results for the inventory dynamic are presented in Table 9 and Table 10. The two tables show empirical results for the subsample where $\Delta|I_t - I_t^d| < 0$ and

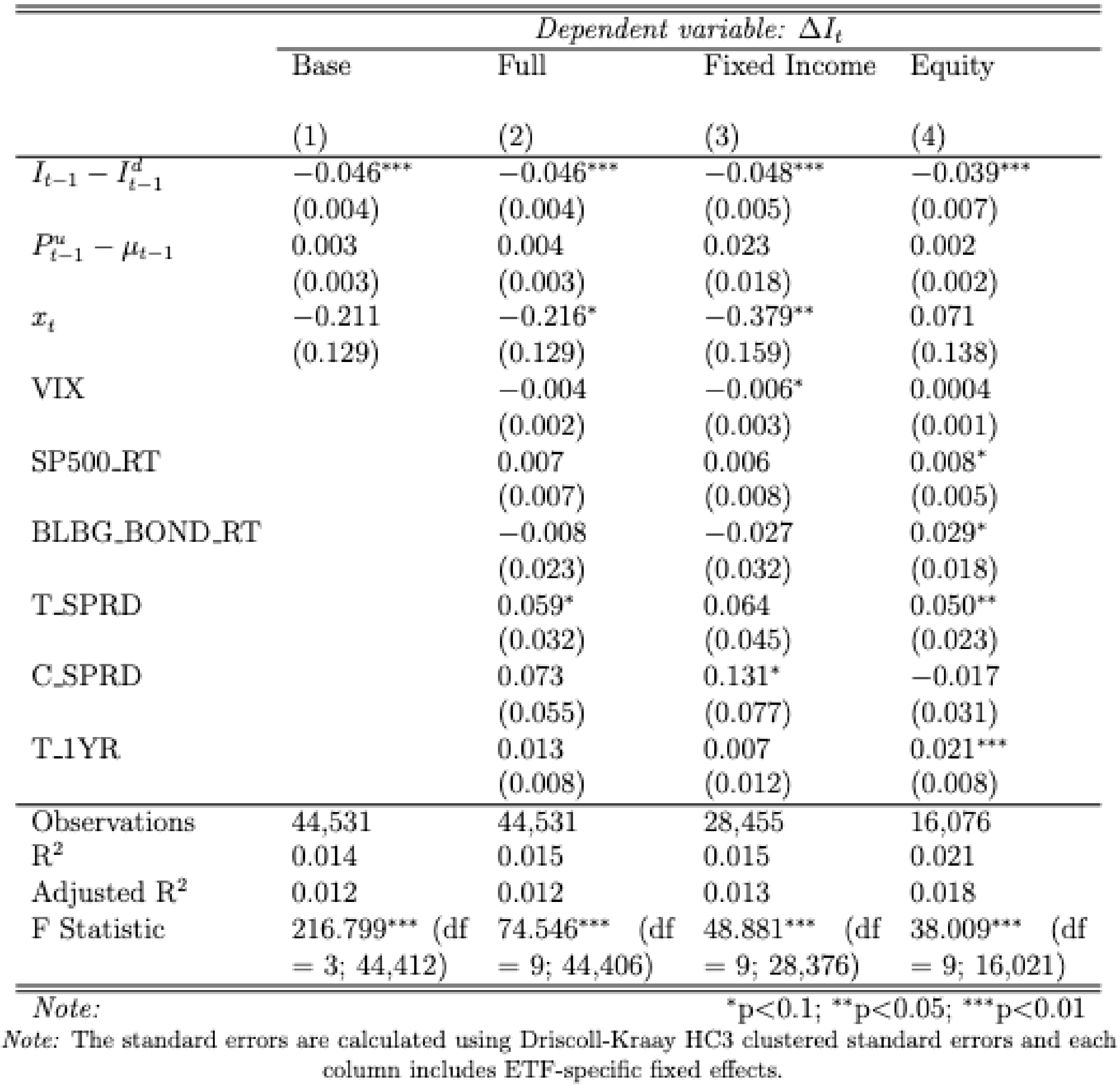
$\Delta\left|I_t - I_t^d\right| > 0$. The former implies that the current period's inventory level is moving towards the optimal level, while the latter suggests that the current period's deviation (in absolute terms) from the optimal inventory level has increased compared to the previous period.

From our empirical exploration, we find that heterogeneous effects exist for these two scenarios. When the APs' inventory is moving closer to the optimal level, inventory management concerns dominate. As shown in Table 9, the previous period's deviation from optimal inventory has a significant negative effect on the current period's inventory change. In contrast, when the current period's inventory is moving further away from the optimal level (Table 10), the previous period's inventory deviation tends to signal the directional position the AP is taking, and this position is likely to persist in the current period as well.

Additionally, the distance between the NAV and fundamental value ($P^u - \mu$) represents an arbitrage opportunity. The APs' current inventory only responds to this in the directional position subsample (though this is significant only in the Equity ETF subsample, possibly due to more costly arbitrage in Fixed Income markets).

The effect of unexpected order imbalance is significant only for Fixed Income ETFs in the inventory management sample, suggesting that these transactions are more representative of a market maker providing liquidity to the market.

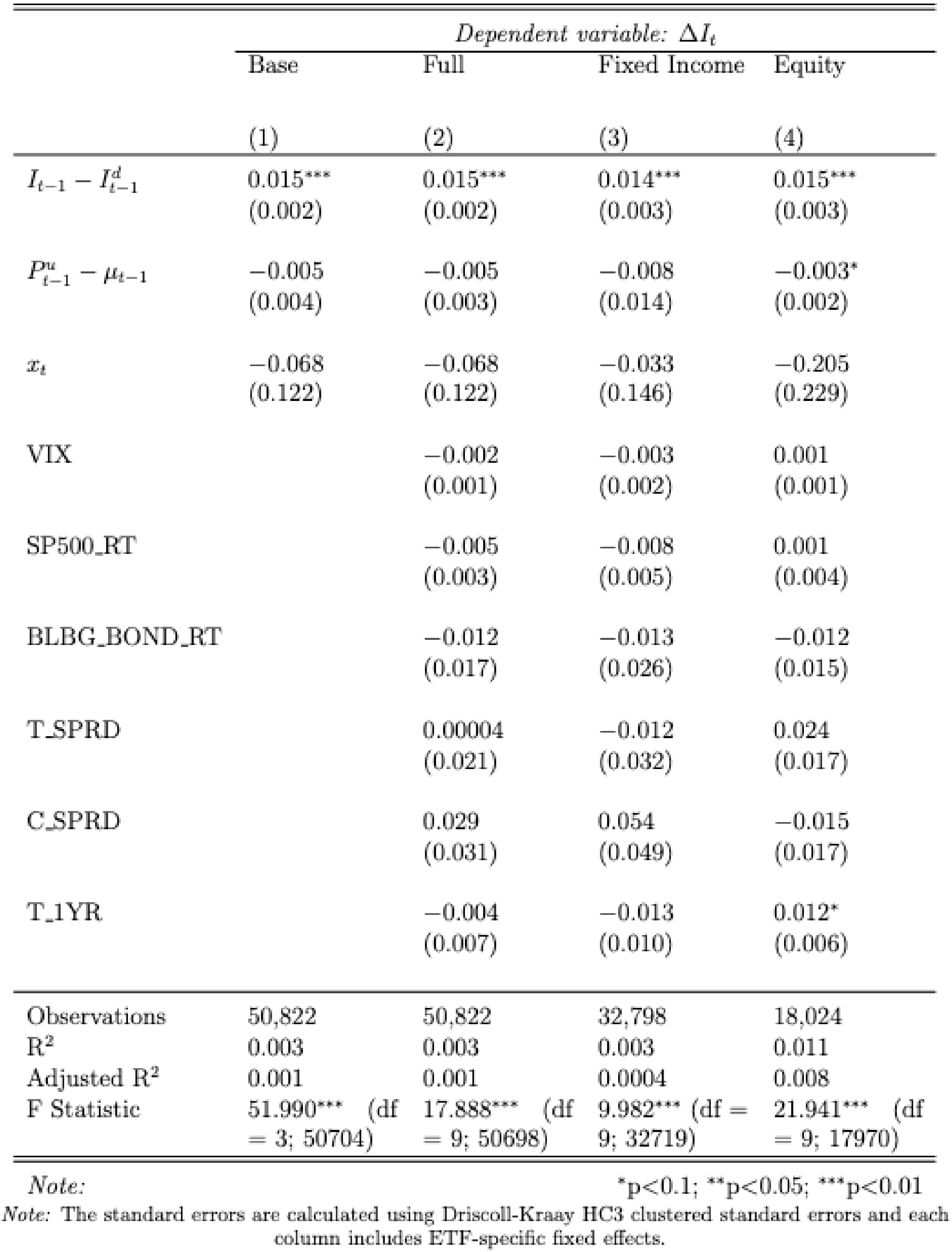
Table 9: Inventory Dynamics (Inventory management subsample)



	Dependent variable: ΔI_t			
	Base	Full	Fixed Income	Equity
	(1)	(2)	(3)	(4)
$I_{t-1} - I^d_{t-1}$	-0.046^{***}	-0.046^{***}	-0.048^{***}	-0.039^{***}
	(0.004)	(0.004)	(0.005)	(0.007)
$P^u_{t-1} - \mu_{t-1}$	0.003	0.004	0.023	0.002
	(0.003)	(0.003)	(0.018)	(0.002)
x_t	-0.211	-0.216^{*}	-0.379^{**}	0.071
	(0.129)	(0.129)	(0.159)	(0.138)
VIX		-0.004	-0.006^{*}	0.0004
		(0.002)	(0.003)	(0.001)
SP500_RT		0.007	0.006	0.008^{*}
		(0.007)	(0.008)	(0.005)
BLBG_BOND_RT		-0.008	-0.027	0.029^{*}
		(0.023)	(0.032)	(0.018)
T_SPRD		0.059^{*}	0.064	0.050^{**}
		(0.032)	(0.045)	(0.023)
C_SPRD		0.073	0.131^{*}	-0.017
		(0.055)	(0.077)	(0.031)
T_1YR		0.013	0.007	0.021^{***}
		(0.008)	(0.012)	(0.008)
Observations	44,531	44,531	28,455	16,076
R^2	0.014	0.015	0.015	0.021
Adjusted R^2	0.012	0.012	0.013	0.018
F Statistic	216.799^{***} (df = 3; 44,412)	74.546^{***} (df = 9; 44,406)	48.881^{***} (df = 9; 28,376)	38.009^{***} (df = 9; 16,021)
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: The standard errors are calculated using Driscoll-Kraay HC3 clustered standard errors and each column includes ETF-specific fixed effects.

Table 10: Inventory Dynamics (Directional position subsample)



	Dependent variable: ΔI_t			
	Base	Full	Fixed Income	Equity
	(1)	(2)	(3)	(4)
$I_{t-1} - I^d_{t-1}$	0.015***	0.015***	0.014***	0.015***
	(0.002)	(0.002)	(0.003)	(0.003)
$P^u_{t-1} - \mu_{t-1}$	−0.005	−0.005	−0.008	−0.003*
	(0.004)	(0.003)	(0.014)	(0.002)
x_t	−0.068	−0.068	−0.033	−0.205
	(0.122)	(0.122)	(0.146)	(0.229)
VIX		−0.002	−0.003	0.001
		(0.001)	(0.002)	(0.001)
SP500_RT		−0.005	−0.008	0.001
		(0.003)	(0.005)	(0.004)
BLBG_BOND_RT		−0.012	−0.013	−0.012
		(0.017)	(0.026)	(0.015)
T_SPRD		0.00004	−0.012	0.024
		(0.021)	(0.032)	(0.017)
C_SPRD		0.029	0.054	−0.015
		(0.031)	(0.049)	(0.017)
T_1YR		−0.004	−0.013	0.012*
		(0.007)	(0.010)	(0.006)
Observations	50,822	50,822	32,798	18,024
R^2	0.003	0.003	0.003	0.011
Adjusted R^2	0.001	0.001	0.0004	0.008
F Statistic	51.990*** (df = 3; 50704)	17.888*** (df = 9; 50698)	9.982*** (df = 9; 32719)	21.941*** (df = 9; 17970)
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: The standard errors are calculated using Driscoll-Kraay HC3 clustered standard errors and each column includes ETF-specific fixed effects.

7 Conclusion

In this paper, we explored the relationship between ETF mispricing and the inventory management practices of Authorised Participants (APs). Using a novel dataset that includes both primary and secondary market data for 128 ETFs over a 5-year period, we were able to directly observe the inventory levels of individual APs and their effects on ETF mispricing. Our findings suggest that APs play a dual role as market makers and arbitrageurs, and that their inventory constraints, especially during periods of market stress, can significantly influence ETF pricing dynamics.

Our dynamic model indicates that APs' pricing and inventory management decisions are driven by a combination of market-making obligations, arbitrage opportunities, and balance sheet constraints. During periods of market stress, such as the COVID-19 pandemic, these constraints become more binding, leading to substantial mispricing in certain ETF categories. The empirical results show that APs' ability to correct mispricing is limited when they approach their inventory capacity, and these limitations could be magnified by the volatility and uncertainty characteristic of financial crises.

These findings have implications for market regulation. First, regulators might consider improving disclosure around APs' inventory holdings during periods of market stress. Providing more frequent information will help market participants understand the underlying drivers of mispricing and make more informed decisions.

Second, there may be a need to reassess the regulatory frameworks governing APs and their market-making roles in ETF markets. Ensuring that APs have sufficient flexibility to manage their inventories effectively during periods of stress will help reduce the severity of mispricing. This might involve easing certain capital or balance sheet constraints in times of crisis or providing APs with additional liquidity support.

Finally, the large mispricing observed in bond ETFs during the COVID-19 pandemic highlights the importance of incorporating ETF market dynamics into financial stability frameworks. The establishment of the Federal Reserve's Secondary Market Corporate Credit Facility (SMCCF) in March 2020 recognised the systemic role of ETFs in modern financial markets. Future policy interventions may benefit from being more pre-emptive, addressing potential liquidity and mispricing issues in ETF markets before they escalate into broader financial stability concerns.

In conclusion, while ETFs offer investors an efficient means of gaining exposure to a wide range of assets, their pricing dynamics are closely linked to the behaviour and constraints of APs. As our results show, APs' inventory management practices can exacerbate mispricing, particularly during periods of stress. Addressing these challenges through regulatory adjustments and improved market functioning will help ensure the continued robustness of the ETF market.

8 References

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. Journal of Financial Economics, 17(2):223-249.

Aquilina, M., Croxson, K., Valentini, G. G., and Sun, Z. (2021). Fixed income ETFs: Secondary market participation and resilience during times of stress. FCA Research Note January 2021.

Aquilina, M., Croxson, K., Valentini, G. G., and Vass, L. (2020). Fixed income ETFs: Primary market participation and resilience of liquidity during periods of stress. Economics Letters, 193:109249.

Aramonte, S. and Avalos, F. (2020). The recent distress in corporate bond markets: Cues from ETFs. Technical report, Bank for International Settlements.

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88(421):284-297.

Durbin, J. and Koopman, S. J. (2012). Time series analysis by state space methods, volume 38. OUP Oxford.

European Securities and Markets Authority (2016). Transaction reporting, order record keeping and clock synchronisation under mifid ii. https://www.esma.europa.eu/sites/default/files/library/2016-1452_guidelines_mifid_ii_transaction_reporting. pdf.

Falato, A., Goldstein, I., and Hortaçsu, A. (2021). Financial fragility in the COVID-19 crisis: The case of investment funds in corporate bond markets. Journal of Monetary Economics, 123:35-52.

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics, 14(1):71100.

Gorbatikov, E. and Sikorskaya, T. (2022). Two APs are better than one: ETF mispricing and primary market participation. Available at SSRN 3923503.

Grossman, S. J. and Miller, M. H. (1988). Liquidity and market structure. Journal of Finance, 43(3):617-633.

Harris, L. (1989). A transaction data study of weekly and intraday patterns in stock returns. Journal of Financial Economics, 16(1):99-117.

Hasbrouck, J. (1991). Measuring the information content of stock trades. Journal of Finance, 46(1):179-207.

Ho, T. S. and Stoll, H. R. (1983). The dynamics of dealer markets under competition. Journal of finance, 38(4):1053-1074.

Hoechle, D. (2007). Robust standard errors for panel regressions with cross-sectional dependence. The Stata Journal, 7(3):281-312.

Kirilenko, A., Kyle, A. S., Samadi, M., and Tuzun, T. (2017). The flash crash: High-frequency trading in an electronic market. Journal of Finance, 72(3):967-998.

Koont, N., Ma, Y., Pástor, L., and Zeng, Y. (2022). Steering a ship in illiquid waters: Active management of passive funds. Technical report, National Bureau of Economic Research.

Laipply, S. and Madhavan, A. (2020). Pricing and liquidity of fixed income ETFs in the COVID-19 virus crisis of 2020. The Journal of Index Investing, 11(3):7-19.

Lettau, M. and Madhavan, A. (2018). Exchange-traded funds 101 for economists. Journal of Economic Perspectives, 32(1):135-154.

Madhavan, A. and Smidt, S. (1993). An analysis of changes in specialist inventories and quotations. Journal of Finance, 48(5):1595-1628.

Mankad, S., Michailidis, G., and Kirilenko, A. (2013). Discovering the ecosystem of an electronic financial market with a dynamic machine-learning method. Algorithmic Finance, 2(2):151-165.

Monache, D. D., Petrella, I., and Venditti, F. (2021). Price dividend ratio and long-run stock returns: A score-driven state space model. Journal of Business & Economic Statistics, 39(4):1054-1065.

Raddatz, C. E. (2021). Does the diversity and solvency of authorized participants matter for bond ETF arbitrage? Evidence from the dash for cash episode. Available at SSRN: https://ssrn.com/abstract=3868529.

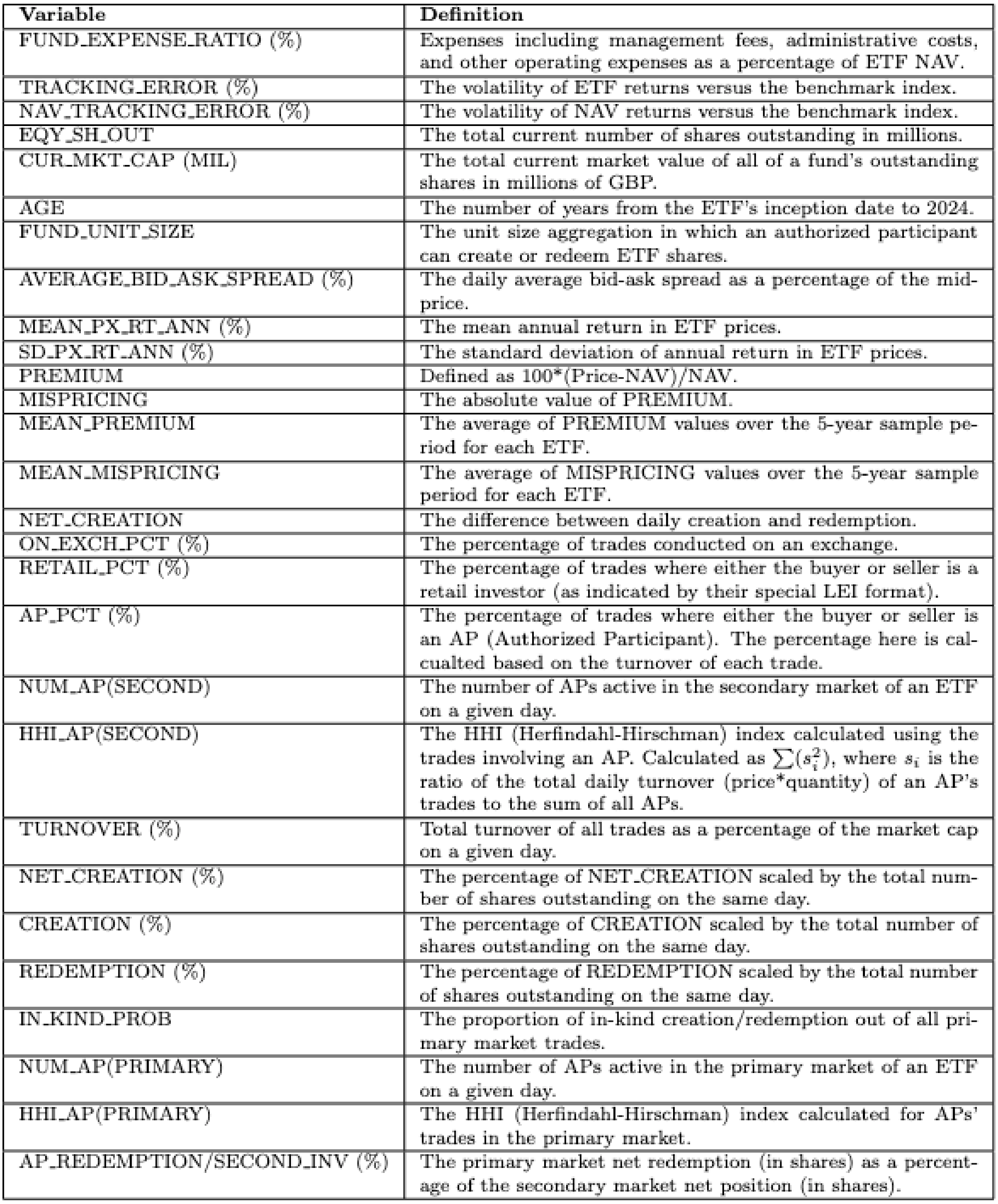
Shen, P. (2002). Market-timing strategies that worked. FRB of Kansas City Research Working Paper No. 02-01, Available at SSRN: https://ssrn.com/abstract=445920.

Shim, J. J. and Todorov, K. (2023). ETFs, Illiquid Assets, and Fire Sales. Available at SSRN: https://ssrn.com/abstract=3886881

Tsay, R. S. (1986). Time series model specification in the presence of outliers. Journal of the American Statistical Association, 81(393):132-141.

Annex 1: Variable Definitions

Table 11: Definitions of variables in the summary statistics table



Variable	Definition
FUND_EXPENSE_RATIO (%)	Expenses including management fees, administrative costs, and other operating expenses as a percentage of ETF NAV.
TRACKING_ERROR (%)	The volatility of ETF returns versus the benchmark index.
NAV_TRACKING_ERROR (%)	The volatility of NAV returns versus the benchmark index.
EQY_SH_OUT	The total current number of shares outstanding in millions.
CUR_MKT_CAP (MIL)	The total current market value of all of a fund's outstanding shares in millions of GBP.
AGE	The number of years from the ETF's inception date to 2024.
FUND_UNIT_SIZE	The unit size aggregation in which an authorized participant can create or redeem ETF shares.
AVERAGE_BID_ASK_SPREAD (%)	The daily average bid-ask spread as a percentage of the mid-price.
MEAN_PX_RT_ANN (%)	The mean annual return in ETF prices.
SD_PX_RT_ANN (%)	The standard deviation of annual return in ETF prices.
PREMIUM	Defined as 100*(Price-NAV)/NAV.
MISPRICING	The absolute value of PREMIUM.
MEAN_PREMIUM	The average of PREMIUM values over the 5-year sample period for each ETF.
MEAN_MISPRICING	The average of MISPRICING values over the 5-year sample period for each ETF.
NET_CREATION	The difference between daily creation and redemption.
ON_EXCH_PCT (%)	The percentage of trades conducted on an exchange.
RETAIL_PCT (%)	The percentage of trades where either the buyer or seller is a retail investor (as indicated by their special LEI format).
AP_PCT (%)	The percentage of trades where either the buyer or seller is an AP (Authorized Participant). The percentage here is calcualted based on the turnover of each trade.
NUM_AP(SECOND)	The number of APs active in the secondary market of an ETF on a given day.
HHI_AP(SECOND)	The HHI (Herfindahl-Hirschman) index calculated using the trades involving an AP. Calculated as $\sum(s_i^2)$, where s_i is the ratio of the total daily turnover (price*quantity) of an AP's trades to the sum of all APs.
TURNOVER (%)	Total turnover of all trades as a percentage of the market cap on a given day.
NET_CREATION (%)	The percentage of NET_CREATION scaled by the total number of shares outstanding on the same day.
CREATION (%)	The percentage of CREATION scaled by the total number of shares outstanding on the same day.
REDEMPTION (%)	The percentage of REDEMPTION scaled by the total number of shares outstanding on the same day.
IN_KIND_PROB	The proportion of in-kind creation/redemption out of all primary market trades.
NUM_AP(PRIMARY)	The number of APs active in the primary market of an ETF on a given day.
HHI_AP(PRIMARY)	The HHI (Herfindahl-Hirschman) index calculated for APs' trades in the primary market.
AP_REDEMPTION/SECOND_INV (%)	The primary market net redemption (in shares) as a percentage of the secondary market net position (in shares).

Annex 2: Data Cleaning

We illustrate below 3 important steps for data cleaning.

Deleting intra trades

The Level II data provided by the Global Legal Entity Identifier Foundation offers information on the ownership structure of legal entities. Each legal entity with an LEI reports their 'direct accounting consolidating parent' and 'ultimate accounting consolidating parent'. This allows us to construct a unique dataset specifying each organization as a concatenated string of LEIs of its members. This allows us to identify if a transaction is an intra trade (i.e., a trade that happens inside an organization). This is typically treated as a practice of internal financial management without the same profit motivations as trades with other organizations. Therefore, we exclude these trades[16] from our sample. Out of our whole sample, these kind of intra trades account for 3.5% of all trades, so they account for only a small portion of all reported market transactions.

Deleting duplicated reports

Duplicated reports emerge when we have a transaction chain (which we define as a trade with one end buyer and one end seller[17] but multiple middle entities who connect the two sides without taking on risks themselves) with multiple entities and each of those with reporting obligations submit a line of report. The difficulty lies in two aspects: (i) there's no unique identifier that links these lines of reports to be the same trade; (ii) each reporting entity only reports the segment of a transaction that is visible to it[18].

To solve the first problem, we group up the reports if they share the same timestamp[19], price, price currency and quantity. This is a fair but still a loose condition and one group of reports may include multiple chains that happen to share the same identifying features, but we are conservative here to ensure that we capture a complete chain (or multiple complete chains). For clarification, we denote this grouping as G_1.

The way to deal with the second problem is to notice that each line of reported trades is essentially a collection of trade segments (defined as two identities and one trade direction) known to the reporting entity. For example, a typical trade report can be decomposed into the following segments: buyer-buyer decision maker, buyer decision maker transmitting buyer, transmitting buyer-executing (reporting) entity, executing

[16] Specifically, these trades include those where the buyer and seller, buyer decision maker and seller, seller decision maker and buyer, or buyer decision maker and seller decision maker belong to the same organization.

[17] This can be relaxed in the case of grouped orders.

[18] For example, when a broker/dealer is trading with an institutional investor on behalf of some clients, the institutional investor can only see the broker/dealer on the opposite side instead of the clients.

[19] We round the timestamps to the nearest seconds even though the data is up to microsecond frequency because MiFID II regulation requires that trades on the same chain should share the same time but subject to different granularity requirements. Only the ones directly facing market on a trading venue needs to be accurate at milliseconds or better, others seconds or better. In order to identify a complete chain, we need to round it to its lowest required frequency.

(reporting) entity-transmitting seller, transmitting seller-seller decision maker, seller decision maker seller. Even though each reporting entity may be 'short-sighted', they would have complete information of entities between the start of their reported chain (buyer) to the end of their reported chain (seller). Therefore, reporting entities on the same chain must have overlapping reported segments[20]. The overlapped reported segments are in fact the duplicated reports and the non-overlapping ones are segments of the same chain that complement each other to trace out the complete chain. Considering the possibility of multiple chains sharing the same identifying features, we further group up rows with the same identifying features based on whether they share any overlapping segments with another row in the group. We call this grouping G_2 and they represent subgroups below the groupings G_1. If a row has no shared segments with all rows in a group, then it's more likely it is a different trade that happen to share the same identifying features as the group. We can then re-construct the chain by combining all segments within these subgroups and eliminating repeated segments. Doing this would eliminate any duplicated reports and map out complete chains from the reported transactions.

Delete intermediary matching entities

We can delete intermediary matching entities using the trade capacity field offered by the MDP data. There are three trading capacities that could be reported by the reporting entity of a transaction: dealing on own account (DEAL), matched principal (MTCH) and any other capacity (AOTC). For our purpose, this shows whether the reporting entity has taken on risks and have potential returns from this transaction and therefore should be considered 'important' or it's only matching orders from both sides (MTCH or AOTC) and should be treated as intermediary matching entities and be deleted from our chains to avoid artificial volumes between end buyers/sellers with these matching entities. The prior case is referring to DEAL trades, where the investment firm action its own proprietary trades or act on its own account to fill clients' order. The latter case refers to MTCH and AOTC trades. MTCH trades denotes the trades where the reporting entity interposes itself between the buyer and seller in a way that's not exposed to any market risks and AOTC trades are also mostly agency trading where the reporting entity is neither the buyer nor the seller. (See European Securities and Markets Authority (2016) for more information.)

To identify the important entities in each group in G_2, we go back to the original transaction reports. If the reporting entity is also the buyer/seller, this is always a DEAL trade, and this buyer/seller needs to be one of the entities we keep. If the buyer/seller doesn't have reporting obligations[21] (for example, a retail trader), then they will be on the edge of the trade chain in G_2 (because they can't intermediate). That makes them the end buyer/seller. Now that we have all end buyers and sellers, we can reconstruct the meaningful trades and delete any intermediary entities.

During the implementation of this method, we find there are 4 more adjustments to make due to the features of the dataset. Firstly, we make the simplifying assumption

[20] Even though they might be given different labels. For example, in a trade segment of A selling to B, it might be reported by A as the executing entity selling to buyer, but reported by B as seller selling to the executing entity. This example comes from a trade pattern like: $Client\ 1 \rightarrow\ A \longrightarrow\ B \longrightarrow\ Client\ 2$.

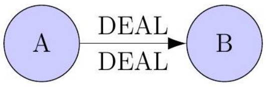
[21] We have information on reporting obligations based on historical transaction records.

that when the same entity appear more than once on buyer/seller side, we only count them as once. This assumes that two parties won't have multiple trades of the exact same identifying features (timestamp, price, currency, quantity) in our group. This is innocuous unless there are lots of iceberg trades. Secondly, our method assumes all entities obligated to report would do so. What we find instead is sometimes when two ends of a transaction report both have reporting obligations, one side doesn't report. This is marked as incomplete reports and account for around 10% of our sample. In this case, we still mark the misreporting side as the end buyer/seller. Theoretically, the misreporting side could be an intermediary linking client side, this would only cause a problem if the client side doesn't have reporting obligation and we mistakenly take the misreporting entity instead of the client as the end buyer/seller. We expect this adjustment to not have a great effect on the analysis as the impact of retail traders are small relatively to institutional investors and this kind of cases should only account for a small part of our sample. Thirdly, we might have some chains in G_2 that have multiple buyers and multiple sellers. When this happens, it's impossible to tell even manually which end buyers are trading with which end sellers (One such case is a trade pattern we call the butterfly trade, see Case 4 below for further discussions on this example). This kind of unidentified case account for around 1% of our sample and in such cases, we randomly allocate end buyers to end sellers. Lastly, there are also grouped trades where an entity helps group up multiple buyer orders and distribute them to multiple seller orders. We don't apply our elimination to this kind of grouped trades because (i) they are not one-to-one relationships even in the actual trades and it would be impossible to eliminate the middle entity (ii) the timing of one side of the INTC trades may span across the whole day (entities which execute grouping trades are required to balance the two sides by the end of the trading day) with multiple executions at different prices and timestamp. The final weighted price would be the price the other side of the clients pay.

We provide some examples of common trade chains we observe from the dataset in the following sections to help with better understanding.

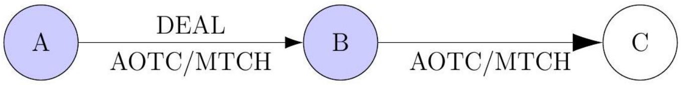
Common trade chains and illustrations of the cleaning strategy

Case 1: Simple DEAL trades



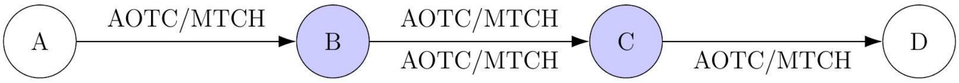
In this case, both A and B are entities with reporting obligations, and we will see two reports of the above segments, and we collapse them into one to eliminate duplicated reports.

Case 2: DEAL + AOTC/MTCH trades



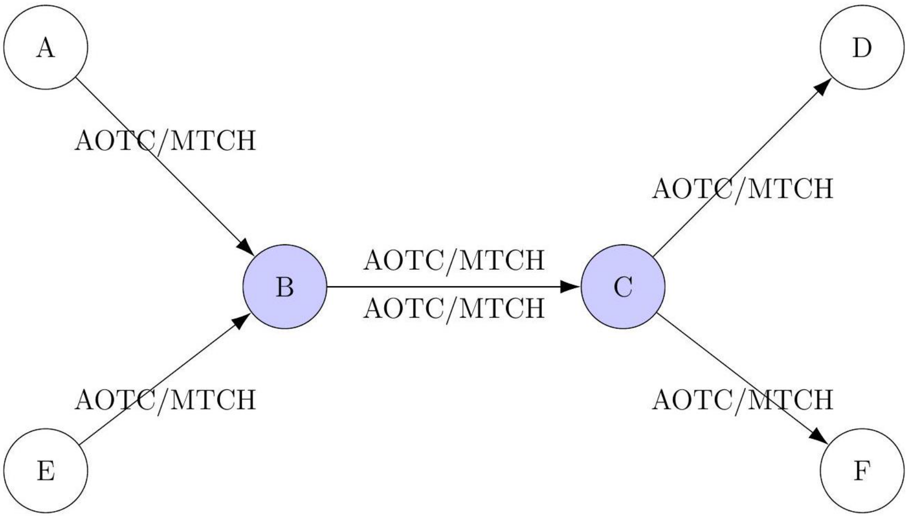
In this case, A and B are entities with reporting obligations and C is a client who doesn't report. B would report an AOTC/MTCH transaction between A and C, but A would report a DEAL transaction with B. After decomposing each line of transaction into segments, we will have two ABs and one BC which we collapse into $A - B - C$. Since B is an entity with reporting obligations but it is not the buyer/seller in its reported transaction, we delete it. A is the seller in its reported transaction, and C is the end buyer, and it has no reporting obligation. Therefore, according to our method, we delete B and connect A and C to be $A - C$ as our final cleaned chain.

Case 3: AOTC/MTCH + AOTC/MTCH trades



In this case, B and C are entities with reporting obligations and A and D are clients who don't report. Breaking each transaction into segments, we would have two reports of BC, one report of AB and CD. According to the rule, we can collapse them into $AB - BC - CD$. Since B and C have reporting obligations and neither of them is the end buyer/seller in their own reports, we delete them and keep A and D, who are the end seller and buyer without reporting obligations. Therefore, we end up with $A - D$ as our final chain.

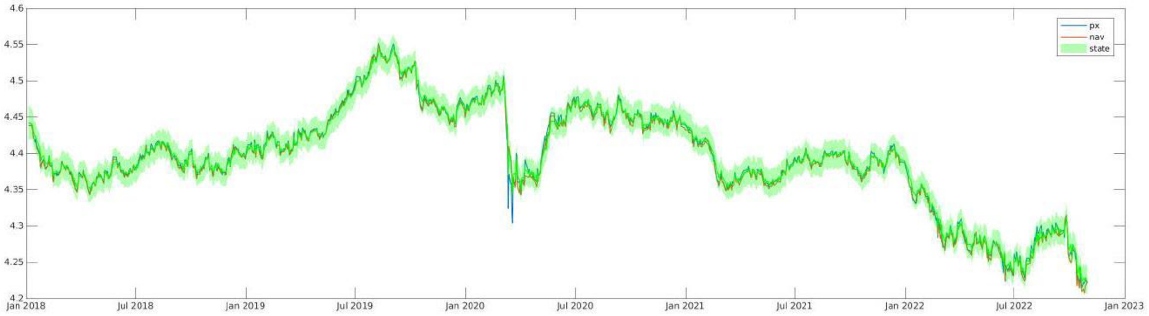
Case 4 (Unidentified): Butterfly trades



In this case, B and C are reporting entities and A, E, D, F are clients who don't report. Breaking each transaction into segments, we would have two BCs, and one $AB, EB, CD, CF.$ Collapsing these segments would delete the extra BC. Because B and C are not the end buyer/seller in their own reports, they are deleted, and we keep all the client end buyer/seller. However, it's hard to know whether A traded with D and E traded with F or the other way around. We only know that there must be two trade chains in this subgroup. In this case, we randomly match the end buyer and seller $A - F, E - D$ or $A - D, E - F$ as the final chain.

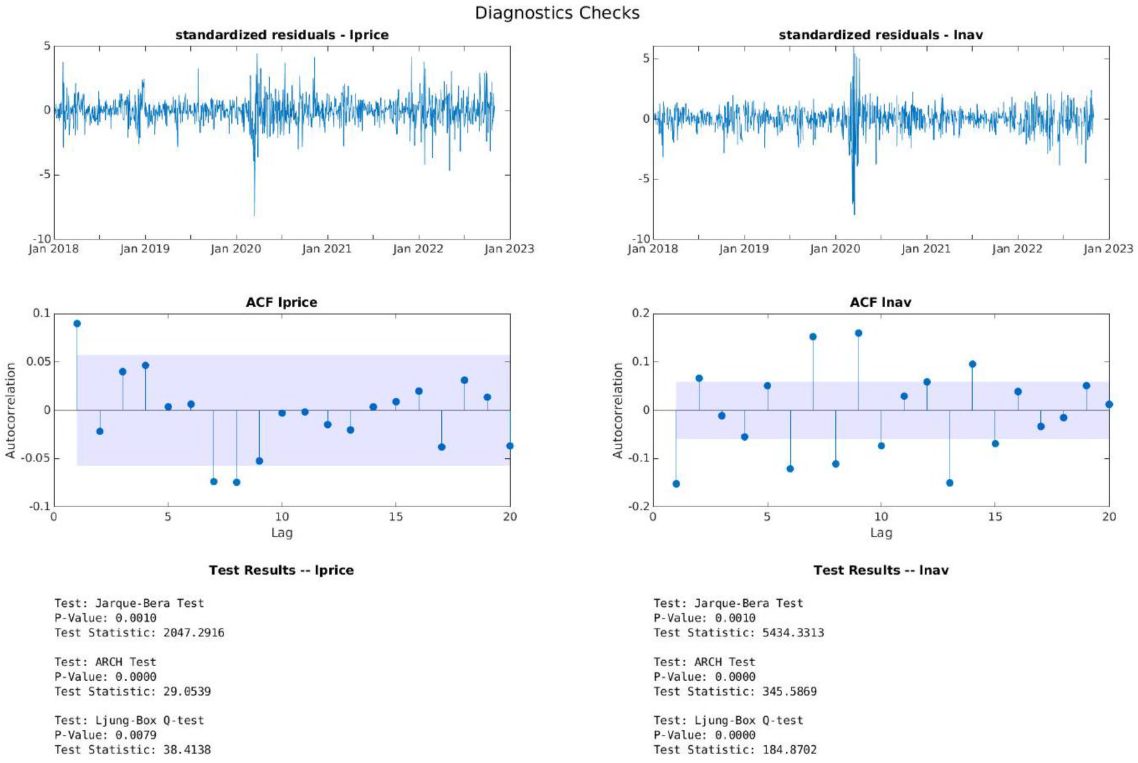
Annex 3: Fitted results for another representative ETF (Baseline: Figure 20 and Score-driven: Figure 21)

Figure 20: Baseline Fitted Model
for iShares J.P. Morgan $ EM Bond UCITS ETF



(a) Price, NAV, Fundamental Value Chart

The above figure demonstrates the ETF price (blue), NAV (red), and the filtered first state variable (bright green) that corresponds to M_t or μ_t. The bands around the green line denote the 95% confidence interval constructed using the state forecast variance for each time period.



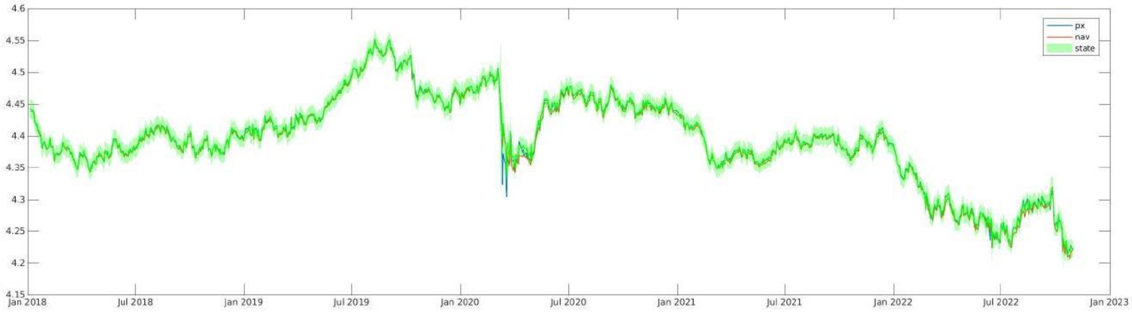
```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 2047.2916

Test: ARCH Test
P-Value: 0.0000
Test Statistic: 29.0539

Test: Ljung-Box Q-test
P-Value: 0.0079
Test Statistic: 38.4138
```

```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 5434.3313

Test: ARCH Test
P-Value: 0.0000
Test Statistic: 345.5869

Test: Ljung-Box Q-test
P-Value: 0.0000
Test Statistic: 184.8702
```

(b) Diagnostics

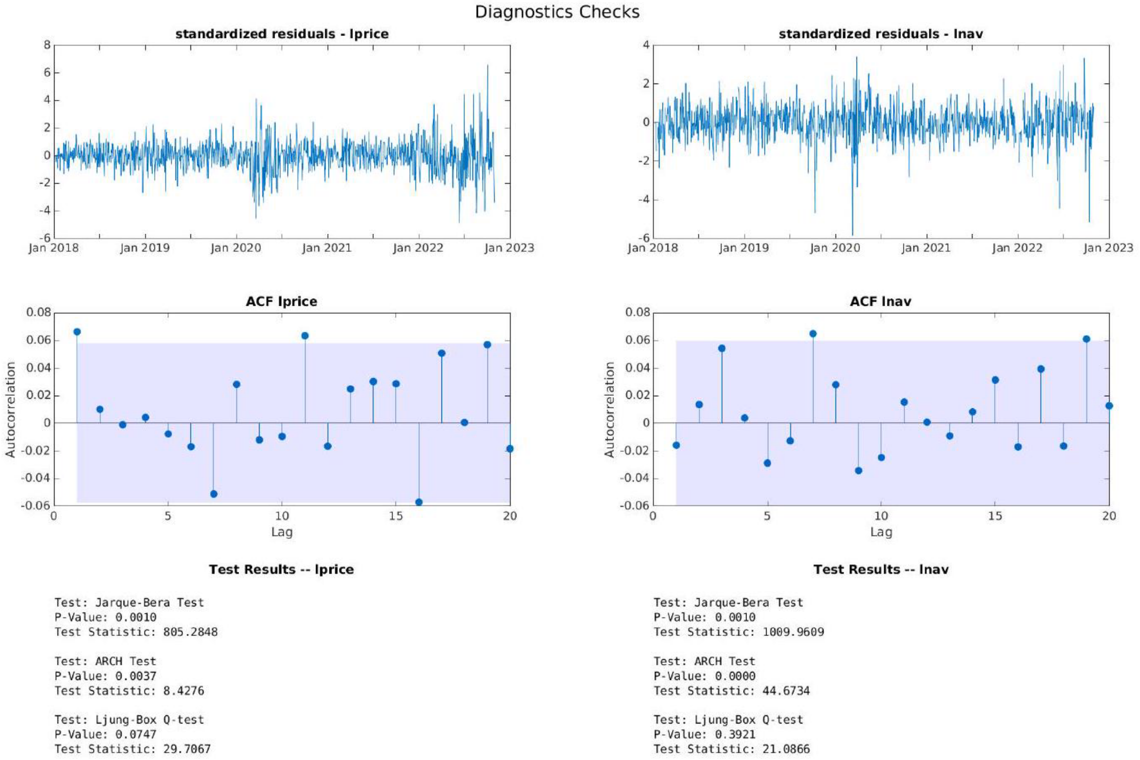
The above demonstrates the diagnostic checks for the baseline state-space model described in paragraph 5.0.1. The first row shows the standardized forecast residuals for predicting the two variables in the measurement equation: $\log$ (price) and $\log$ (NAV). The second row shows the ACF plots for the standardized residuals along with Bartlett's confidence *interval. The third row presents the test statistics and p-values for the two measurement residuals: normality test, heteroskedasticity test, and autocorrelation test.*

Figure 21: Time-varying Parameters Fitted Model for iShares J.P. Morgan $ EM Bond UCITS ETF



(a) Price, NAV, Fundamental Value Chart

The above figure demonstrates the ETF price (blue), NAV (red), and the filtered first state variable (bright green) that corresponds to M_t or μ_t. The bands around the green line denote the 95% confidence interval constructed using the state forecast variance for each time period.



```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 805.2848

Test: ARCH Test
P-Value: 0.0037
Test Statistic: 8.4276

Test: Ljung-Box Q-test
P-Value: 0.0747
Test Statistic: 29.7067
```

```
Test: Jarque-Bera Test
P-Value: 0.0010
Test Statistic: 1009.9609

Test: ARCH Test
P-Value: 0.0000
Test Statistic: 44.6734

Test: Ljung-Box Q-test
P-Value: 0.3921
Test Statistic: 21.0866
```

(b) Diagnostics

The above demonstrates the diagnostic checks for the score-driven time-varying-parameter state-space model. The first row shows the standardized forecast residuals for predicting the two variables in the measurement equation: $log\ (\ price\ )$ and $log\ (NAV)$. The second row shows the ACF plots for the standardized residuals along with Bartlett's confidence interval. The third row presents the test statistics and p-values for the two measurement residuals: normality test, heteroskedasticity test, and autocorrelation test.

FINANCIAL
CONDUCT
AUTHORITY

12 Endeavour Square, London E20 1JN
Telephone: +44 (0)20 7066 1000 Website: www.fca.org.uk