Evolutionary Arbitrage

Julio Crego, Jens Kvaerner, Aavald Sommervoll,

Dag Sommervoll and Niek Stevens*

November 25, 2022

Abstract

The prices of exchange-traded funds (ETFs) can deviate significantly from their net asset values (NAVs). Exploiting such inefficiencies is often too costly because it involves taking positions in hundreds of underlying illiquid securities. We develop a method that identifies a liquid mimicking portfolio that tracks the NAV using only ETFs. Our method combines a genetic algorithm with non-negative least squares. We apply it to the fixed income ETF market. Our long-short strategy generates a Sharpe ratio of 4-5, incurs little transaction cost, and does well under all market conditions.

JEL Codes: G120, G140

*Crego: Tilburg University, Warandelaan 2 5037 AB Tilburg Netherlands; J.A.Crego@tilburguniversity.edu. Kvaerner: Tilburg University, Warandelaan 2 5037 AB Tilburg Netherlands; jkverner@gmail.com. Aavald Sommmervoll: University of Oslo, Postboks 1080, 0316 Oslo, Norway; daavalds@ifi.uio.no. Dag Sommmervoll: Norwegian University of Life Sciences, P.O. Box 5003. NO-1432 Aas, Norway; dag.einar.sommervoll@nmbu.no. Niek Stevens: SIG-i Capital AG; niek.stevens@sig-i-capital.com. We thank Joost Driessen, Alexander Hoffmann and Alvaro Remesal for valuable feedback.

1 Introduction

Financial markets are characterized by multiple securities whose price depends on one or more other securities. Arbitrageurs ensure efficient pricing of such securities by actively looking for mispriced securities and exploiting these opportunities. With unlimited arbitrage capital, arbitrageurs sell overvalued assets and buy undervalued assets until any mispricing disappears (Friedman, 1953). In reality, arbitrage trading is risky and costly, which can cause mispricing to persist (Harrison and Kreps, 1978).

Exchange-traded funds (ETFs) are an example of an asset class whose price depends on the underlying securities. The theoretical price of the ETF is equal to the sum of all its securities, called the net asset value (NAV). An arbitrage opportunity occurs if the ETF price differs from its NAV. In these situations, the so-called authorized participants (AP) can create or redeem shares of the ETF at the NAV. This arbitrage mechanism creates a close link between the ETF and its underlying securities (Ben-David et al., 2018), where the riskiness of arbitrage trading increases in the illiquidity of the underlying securities.

Corporate bond ETFs are one subclass of ETFs considered risky for arbitrage due to the illiquidity of the underlying securities (Petajisto, 2017; Dannhauser and Hoseinzade, 2021). The lack of arbitrage trading reduces market-making activities in this market, which decreases liquidity and can lead to persistent mispricing. The Covid-19 outbreak is one example in which the Federal Reserve intervened as the dealer of last resort.

The fixed income ETF asset class has developed from less than $10 billion in 2009 to more than $1.2 trillion today (Todorov, 2021). Given the size of the asset class and the difficulty in exploiting arbitrage in this market, we need new methods to make it more attractive for arbitrageurs to ensure efficient pricing. In this paper, we develop an empirical strategy that makes it possible to trade on market inefficiencies while circumventing the *de facto* non-tradability of the NAV. Unlike traditional arbitrage, our strategy does not take direct positions in the underlying securities.

The backbone of our approach is to identify a portfolio of liquid ETFs that work as a

substitute for the underlying securities that make up the NAV. Econometrically, we combine natural selection (a genetic algorithm) with nonnegative least squares (NNLS). The economics behind this approach is straightforward: new ETFs will be nearly redundant—both with respect to the underlying securities and outstanding ETFs—as the number of ETFs grows. As a result, arbitrageurs can exploit mispricing by trading the ETF (which can be higher or lower than its NAV) and cover the position with a portfolio of other ETFs. We label the long position the *mimicking portfolio*.

A valuable mimicking portfolio must have several attributes (Roll and Srivastava, 2018). First, it must consist only of liquid assets tradable at low cost. Second, a regression of the target price on the mimicking portfolio must yield a sufficiently high R^2. In other words, the mimicking portfolio must explain most of the price variation of the target. Finally, it must be possible to trade in real-time. Our mimicking portfolio satisfies these requirements.

A set of constraints ensures that mispricing is exploitable at low cost and for a range of volumes. First, the mimicking portfolio contains, at most, eight ETFs. Second, nonnegative least squares ensure that all weights in the mimicking portfolio are positive. Third, we always sell the target in whole units. Last, to make it also work for low volumes, we allow the weights in the mimicking portfolio to have up to two decimals.

We control for data snooping in two ways. First, following standard practice in machine learning with time-series data, we separate between a training period and a test period. We use the training period to identify the genetic footprint of the potential (mispriced) targets and the number of assets in the mimicking portfolio. We use the test period, which starts after the training period, for performance evaluation.

While the OLS estimator finds the weights in the mimicking portfolio that maximize the portfolio's ability to explain price variation in the target, it breaks down in practice. The reason is the high collinearity of the regressors. This multicollinearity causes a large positive coefficient of one regressor to be offset by a large negative coefficient of another highly correlated regressor: each (large) position in the mimicking portfolio is offset by a

(large) short position. The nonnegative least squares estimator prevents such instability.

Our empirical tests start in 2010. We hold all positions until the next trading day. Our annual average daily returns are between 1 percent (2010) and 56.9 percent (2020). The corresponding standard deviations are between 1.6 percent (2017) and 11.8 percent (2010). From 2013, the mean return is at least four times greater than the standard deviation. The annualized Sharpe ratios are between 0.1 (2010) and 8.9 (2015), with a median of 5.3. After the transaction cost, the estimated Sharpe ratio drops to about 3.

We shed some light on the drivers of arbitrage profits. First, we investigate whether exposure to common equity and bond factors explains the return on the arbitrage strategy. The portfolio's loadings on all factors are small and statistically insignificant: the return on the arbitrage portfolio is market neutral. Second, we relate the variation in arbitrage profits to the variation in the implementation costs and the riskiness of the trades. High profits are associated with: i) large spreads between the ETF and the NAV, often occurring during financial market turmoil (see e.g., Madhavan and Sobczyk, 2016), and ii) risky long-short positions; either in the form of a noisy mimicking portfolio, or an illiquid short position.

We unravel the benefits of each part of our method. Nonnegative least squares is crucial for finding a cheap and reliable mimicking portfolio. Relying on nonnegative least squares, we compare our approach to an alternative that uses only one ETF to hedge the short position. The analysis shows that diversification is essential for arbitrage profits. In the last specification, we introduce diversification but ignore the genetic algorithm. We do so by selecting the eight ETFs with the highest correlation with the target. The central insight is that the genetic algorithm protects profits from crashes during financial market turmoil.

This paper adds to a growing literature that applies machine learning techniques to solve practical problems in finance (see e.g., Ban et al., 2018; Feng et al., 2020; Easley et al., 2021; Brogaard and Zareei, 2021, the latter includes a nice application of evolutionary genetic algorithms). Our specific contribution is to develop a methodology that allow in principle anyone to construct arbitrage strategies at a low cost when the underlying assets

are expensive to trade. The procedure does not require detailed information about the underlying investor clientele. Thus, we do not take a stand on what kind of institutions and private individuals hold and trade the underlying securities and ETFs; we focus solely on the spread created by discontinuously liquid underlying assets and dislocations in the market. We remain close to the literature on optimizing portfolios but with a different objective: to minimize the variance while selling an overvalued asset.

The economics behind the results and the inference we draw from the correlations between arbitrage profits and market conditions are based mainly on the previous literature. For example, Laipply and Madhavan (2020) find that most underlying securities in corporate bond ETFs are illiquid; they estimate that 65-75 percent of the underlying constituents do not trade on a given day. Tucker and Laipply (2013) conclude: "Broadly speaking, ETFs that track more-liquid markets (U.S. Treasures and investment-grade credit) tend to have shorter half-lives than those that track less-liquid markets (municipals and high yields)."

With granular data on holdings, primarily available from 2017 (see Shim and Todorov, 2022), one could shed more direct light on the economics behind our results—for example to what extent the profits from the arbitrage trades are predictable by the degree of illiquidity in the basket of underlying securities. In addition, one could, in principle, also estimate a demand system in which one studies counterfactual scenarios to understand how large scale use of our approach would impact the robustness of the overall bond market.

2 Methodology

In this section, we first explain the market structure for exchange-traded funds (ETFs) and how arbitrageurs improve market efficiency. We highlight a limitation with traditional arbitrage for contingent claims whose value depends on illiquid securities. We then develop a method to exploit mispriced contingent claims whose price can depend on an unlimited set of illiquid securities. Our method does not use the underlying securities and hence serves as

an alternative to market-driven inventory management.

2.1 Traditional ETF Arbirtrage

Exchange-traded funds (ETFs) are contingent claims. Contingent claims are securities whose price depends on the price of one or more other securities. ETFs contain a basket of securities, such as stocks and bonds. The ETF price equals the price of all the securities in the fund referred to as the Net Asset Value (NAV).

Like mutual funds, ETFs allow investors to buy a basket of securities in the secondary market (an exchange). Unlike mutual funds, ETF shares can be created or redeemed in the primary market. The primary ETF market consists of the ETF sponsor (e.g., Blackrock, ProShares, Vanguard) and an Authorized Participant (AP). The AP is usually a large financial institution. The ETF sponsor enters a legal contract with the AP, which in turn interacts directly with the financial market.

New ETF shares are created when an AP deposits a basket of underlying securities. The AP can hold the new ETF shares or sell them on the secondary market. Share redemption works the other way around. The AP redeems shares by transferring ETF shares to the ETF sponsor in exchange for the basket of underlying securities. Share creation and redemption keep the ETF price close to its NAV.

An arbitrage opportunity occurs if the ETF price differs from its NAV. Arbitrageurs eliminate these opportunities by relying on market-driven inventory management. For example, during periods with high demand for buying a particular ETF, its price can rise above its NAV. At this point, the APs come into play. They can purchase the underlying basket, deliver it to the provider in exchange for new ETF shares, and then sell the obtained shares in the market again. This operation would produce an arbitrage profit and push the ETF price back towards its NAV. Similarly, during periods with high demand for selling a particular ETF, the arbitrage process works in the same way, but in the opposite direction.

Improving market efficiency through market-driven inventory management does not work

well for all contingent claims. Mispriced claims on illiquid securities are risky and costly to eliminate, which can cause mispricing to persist (Harrison and Kreps, 1978). Fixed income ETFs are a good example of such claims. Its underlying securities are mainly traded over-the-counter (OTC). Securities traded OTC are less liquidity and transparent than products traded on a primary exchange. Spreads in OTC markets tend to be quite wide and thus the OTC market can be opaque and discontinuously liquid (Bessembinder and Maxwell, 2008).

2.2 Alternative ETF Arbitrage

Our arbitrage trading methodology involves the following steps. First, for each day, we search the universe of ETFs to find a target. We always sell the target, thus restricting the units short to integers.[1] We then fit non-negative least squares (NNLS) using the target price as the dependent variable and up to 8 ETFs as independent variables. The weights in the mimicking portfolio are the coefficients in this regression. For the mimicking portfolio, which by definition consists only of long positions, we allow positions up to two decimals.[2] These restrictions ensure that the trading strategy we propose has low transaction costs and can be traded with minimal price impact.

2.2.1 The Target

We identify a target to short as follows. From the first trading date and onward, we select all ETFs which have been traded every day over the last 500 days. These are the candidate targets. All candidate targets are ranked based on their ETF premium ($\ln(\text{Price}_{i,t}^{ETF}/\text{NAV}_{i,t})$). We then select randomly one ETF among the three targets with the highest ETF premium. Randomly selecting one of the targets with high price relative to fundamentals ensures that our results come from a diverse set of ETFs.

[1] As a result, we always look for ETFs that trade above their NAV. We do not impose any minimum spread condition on the trading, and the same ETF might be used to replicate multiple targets.

[2] If you would sell 100 contracts then all the weights in the mimicking portfolio are integers. Some brokers allow fractional trading of ETFs which would result in greater flexibility of the developed methodology.

2.2.2 The Mimicking Portfolio

We use a genetic algorithm to identify the best ETFs to include in the mimicking portfolio. The mimicking portfolio plays the role of the underlying securities in traditional arbitrage based on market-driven inventory management. It consists of up to 8 ETFs that were themselves possible targets (i.e., we impose the same data requirements on the ETFs in the mimicking portfolio as for the target). In the following, we explain our methodology. Appendix A contains a discussion of our choice of hyperparameters.

The intuition regarding the construction of a mimicking portfolio is straightforward. As we plan to short an ETF with a high price relative to its NAV, we seek a weighted sum of ETFs that historically mirrors the close price of the target ETF. However, the definition and search for good mimicking portfolios entails several challenges, starting by the criteria to assess these portfolios.

A natural first choice is to evaluate a candidate mimicking portfolio by the fit (measured by R^2) of an OLS regression:

$$\text{Price}_{i,t}^{ETF} = \sum_{j \neq i} \beta_j \text{Price}_{j,t}^{ETF} + \epsilon_{i,t}. \tag{1}$$

This type of unconstrained OLS estimator has several shortcomings due to the high collinearity of regressors. The high correlation tends to give a spurious good fit as it allows a large positive coefficient of one regressor to be offset by a large negative coefficient of another highly correlated regressor. A negative coefficient in a mimicking portfolio represents a short position. In other words, the high degree of collinearity tends to give us a portfolio where we take a large position in some ETFs and balance this with a large short position in other ETFs. Thus, the unconstrained OLS estimator gives unstable portfolio weights that translate into expensive portfolio strategies.

We resolve this issue with nonnegative least squares (NNLS). The NNLS estimator gives stable portfolio weights despite highly correlated regressors, as all ETFs are required to

have a positive contribution to the mimicking portfolio. It also ensures that the mimicking portfolio consists of only long positions.

We fit the NNLS without a constant term. The reason is that we seek a weighted sum of ETFs that jointly mirrors the target price. We use the R^2 specified in equation 2 to evaluate the fit of a candidate mimicking portfolio:

$$R^2 = 1 - \frac{\sum_i (\text{Price}_{i,t}^{ETF} - \widehat{\text{Price}_{i,t}^{ETF}})^2}{\sum_i (\text{Price}_{i,t}^{ETF})^2}, \tag{2}$$

where $\widehat{\text{Price}_{i,t}^{ETF}}$ is the fitted value ETF price.

The R^2 measure for models without a constant term can be higher than a model with a constant term as we divide by $(\sum_i \text{Price}_{i,t}^{ETF})^2$ in contrast to $\sum_i (\text{Price}_{i,t}^{ETF} - \overline{\text{Price}_{i,t}^{ETF}})^2$, with $\overline{\text{Price}_{i,t}^{ETF}}$ denotes the sample mean. The R^2 measure for models without a constant term can also be negative. It is, however, bounded from above at 1 and 1 corresponds to a model that explains all observed variation. We truncate values at 0, so that reported R^2's are in $[0, 1]$, as in the case of R^2's from regressions with constant term.[3]

The next step is to evaluate a large number of mimicking portfolios. To highlight how demanding this is, suppose that we have 100 ETFs and aim to single out one subset as the mimicking portfolio. The number of subsets (the Bell number) is $1.85 \cdot 10^{47}$. If we restrict ourselves to subsets of 8 ETFs (as we do in our analysis), the number is $1.86 \cdot 10^{11}$. The search for good mimicking portfolios is not only challenging due to a large number of possibilities, but the set of candidates also has a complex combinatorial structure. This stems from the trivial observation that a subset of ETFs may have modest replicating power, but combined with another ETF, or two, it can have high replicating power. Such data structures pose challenges for conventional optimizers, which get trapped in local maxima. We solve this problem with a genetic algorithm.

Natural selection works by selecting the most fit individuals for survival and reproduction.

[3] We view a negative R^2 as none of the variation is explained by the model in question.

A genetic algorithm works in the same way. At the high level, we randomly draw 100 subsets of 8 ETFs. We refer to a subset of 8 ETFs as an *individual* and the set of 100 such individuals as the *population*. Each such subset gives rise to the NNLS-regression that identifies the portfolio weights in the mimicking portfolio. We use natural selection to improve the fitness of mimicking portfolio. Fitness measures R^2 and is give by equation 2.

The key ingredient is to use the most fit individuals as a basis for a new population, which we refer to as the next *generation*. The least fit ones are discarded and replaced by the recombinations of the most fit (offspring). The next generation tends to have higher mean fitness, gradient ascent commence. Note that the forementioned strength of the genetic algorithm relies on the possibility for multiple gradient ascents at the same time. Thus, if some gradient ascents may be trapped at local minima, others may still be ascending.

At the detailed level, we first select the 33 most fit individuals and pair those with the group of the next 33, that is, the 34 to 66 most fit individuals. We form 33 pairs by pairing after fitness. In other words, the most fit in the first group is paired with the most fit in the second group. Each pair will serve as parent for an offspring. The offspring is created by a random draw of 4 ETFs from parent 1 and the remaining 4 are drawn from parent 2 ETFs not previously drawn from parent 1. This ensures that the new individual has 8 distinct ETFs.[4] We allow for mutations. The mutation rate, that is the probability that any given ETF in the offspring is replaced by a random draw of an ETF from the set of all possible ETFs, is $p = 1 - \frac{1}{\sqrt[8]{2}} \approx 0.083$. This probability is chosen to obtain a probability of 50 percent that the offspring has at least one mutation. The mutations play an important role in genetic algorithms as genetic variation is crucial for not getting stuck at local maximums. In our case, the importance is easily illustrated by the contingency of some ETF that boosts R^2 is not present in our population. A "fortunate" mutation may then (re)introduce this "gene" into the population and contribute to higher fitness and (gradient) ascent.

Figure 1 illustrates how recombination and mutation work. In the figure, we represent

[4]The technical term for this recombination is so-called genetic crossover.

every ETF by a unique integer. We see that red are ETFs (genes) inherited from Parent one, and blue are ETFs inherited from Parent two. Moreover, one mutation introduced a new ETF by replacing ETF 24 by ETF 53.

Figure 1 Recombination and Mutation

Parent one:	(2 11 13 17 22 23 26 31)
Parent two:	(1 13 16 18 19 23 24 25)
Offspring:	(1 13 16 17 23 24 25 31)
Offspring with mutation:	(1 13 16 17 23 53 25 31)

The 33 pairs of individuals (parent one and parent two) described above give rise to 33 new individuals. These individuals replace the 33 least fit individuals in the population. For this new generation, we again pair the 33 most fit individuals (measured by R^2) with the 34 to 66 most fit individuals, and form 33 new individuals. These individuals replace the 33 least fit individuals again, and the next generation is born. We use 500 generations as a stopping criteria for the algorithm.

The strength of a genetic algorithm is that it combines random draws with non-random selection. The random draws ensure that any subset of 8 ETFs might be chosen. In this respect, the first generation could be viewed as a random search. However, the nonrandom selection ensures that we combine and mutate ETFs subsets of increasingly higher R^2s. This makes the genetic algorithm a gradient ascent algorithm, but with an important twist: there are multiple ascents working in parallel. In consequence, the algorithm is more robust as one ascent may be stuck at a local maximum without stopping the algorithm. In parallel, the algorithm explores different areas of the candidate space. This is important in our case because the set of candidates has a complex combinatorial structure and the number of candidates is high.

Figure 2 illustrates how the nonrandom selection gives more fit individuals from generation to generation. Fitness is measured by the mean absolute value deviation measured in

the basis points of the best mimicking portfolio. We see a rapid improvement during the first 100 generations and more marginal improvement in the remaining 400.

[Insert Figure 2 here]

2.2.3 Profit Calculations

We hold all positions until the next trading day. The profit of the arbitrage portfolio is the price change of the mimicking portfolio minus the price change of the target. Specifically, let $\text{Price}_{S,t}^{ETF}$ be the close price at day t of the target ETF that we short. Let V_t be the value of the mimicking portfolio that we buy on the same day. The mimicking portfolio is a weighted average of ETFs at time t, i.e., $V_t = \sum_i^8 w_{i,t}\text{Price}_{i,t}^{ETF}$ with $w_{i,t} \geq 0$. We normalize the investment in the mimicking portfolio each day so that the price of the arbitrage portfolio is zero. The normalization is small. In practice, the remainder can be financed with short-term borrowing to ensure that all the weights ($w_{i,t}$) in the mimicking portfolio are in two decimals. With no more than two decimals, the arbitrage portfolio can be traded with only integer positions by trading at least 100 contracts. We calculate our main variable of interest, the daily return on the abitrage portfolio as:

$$r_{t+1} = \frac{(V_{t+1} - V_t) - (\text{Price}_{S,t+1}^{ETF} - \text{Price}_{S,t}^{ETF})}{\text{Price}_{S,t}^{ETF}} = \frac{(V_{t+1} - \text{Price}_{S,t+1}^{ETF})}{\text{Price}_{S,t}^{ETF}}. \tag{3}$$

3 Empirical Tests

In this section, we apply the arbitrage strategy developed in the previous section to the fixed income ETF market. The section proceeds as follows. First, we describe our dataset and present descriptive statistics. We then present the main results focusing on: i) the Sharpe ratios of the arbitrage strategy, and ii) transaction costs. We then analyze whether the returns of the arbitrage strategy can be explained by common risk factors in the bond and equity markets. The last parts focus on the drivers of arbitrage profits and the importance

of combining NNLS with a genetic algorithm for the stability of the results.

3.1 Data

3.1.1 Data sources and sample selection

Our data includes ETFs traded between 2002 and 2020. The data frequency is daily. We construct our sample as follows. First, we select all funds with non-missing 8 digits CUSIP (cusip8) and Lipper-class definition (Lipper_class), classified as ETFs (et_flag equal to F). We include ETFs defined as Corporate, Government, Money Market, and Domestic Equity, and group these into three asset classes: corporate credit (C), government and money market funds (G), and US equity (E). Table 3 presents the corresponding Lipper classifications.

From the monthly CRSP file, we obtain the ETF price (PRC), denoted by ETF $\text{Price}_{i,t}$, volume (VOL), and 8 digits CUSIP for all securities with share code (SHRCD) equal to 73. We then take the intersection of these two datasets, which we merge with NAV (dnav). We exclude ETFs from our sample that have an ETF premium, defined as $\ln(\text{Price}_{i,t}^{ETF}/\text{NAV}_{i,t})$, above 15%. Such large deviations are unlikely to reflect mispricing, but rather measurement errors. Overall, these filters result in 859 unique ETFs and about 1.17 million fund-day observations. We refer to this as our base sample. The number of unique ETFs in the base sample increases drastically over our sample period: from 55 in 2002 to 675 in 2020. We only search for mispricing among ETFs with the Lipper-class definition Corporate (C). From the base sample, we create a dynamic sample recursively based on the following decision rules. The algorithm requires approximately two years of data, i.e., 500 days, which means that we implement the first trade at the beginning of 2002. If an ETF is removed, we set the ETF price equal to the last day before removal.

3.1.2 Descriptive statistics

Table 4 presents descriptive statistics for the number of unique ETFs by asset class for the full sample period that satisfy our filter. In 2006, one fixed income ETF satisfies our filter,

which increases to seven in 2008 and further to 20 in 2010. In the same years, 118, 248, and 264 equity ETFs satisfy the same criteria. Because we need a fairly large set of ETFs to choose from to find a good mimicking portfolio, we start in 2010.

[Insert Table 4 here]

Table 5 presents descriptive statistics for volume and bid-ask spreads by year and asset class from 2010 to 2020.

[Insert Table 5 here]

The upper part of Table 5 reports average daily trading volume by year and asset class. The average number of shares traded within a day for the median corporate credit ETF is 17.4 thousand in 2010 and 36.4 thousand in 2020. The daily volume is about the same for corporate credit as for equity. The average number of shares traded of Government and Money Market ETFs is similar to the one for credit and equity in the early 2010s but increases substantially towards the end of the sample period. The left panel reports the average daily trading volume for ETFs in the 10th percentile. The right panel reports the corresponding statistics for the 90th percentile. The volume distribution is extremely skewed for all asset classes. For corporate credit, the average daily trading volume for ETFs in the bottom 10 percent of the distribution is less than 1 thousand. In contrast, it is typically traded between 300 and 400 thousand contracts each day in the top 10 percent of the distribution.

The lower part of Table 5 reports the bid-ask spread scaled by the closing price by year and asset class multiplied by 100. The median spread for corporate credit and government and money market ETFs is between 0.2 and 0.4 during the sample period. Equity ETFs have larger spreads. The left and right panels report the spreads for ETFs in the 10th and 90th percentiles of the spread distribution. For corporate credit, the spreads are roughly between 0.1 and 1. The spread interval corresponds roughly to an Ask/Bid ratio between 40 and 400 basis points, with the median between 80 and 160 basis points.

The spread intervals are calculated as follows. Assume that the closing price is equal to the average of the ask and the bid price. It then follows that $(\text{Ask}-\text{Bid})/(0.5\times(\text{Ask}+\text{Bid}) = x$ (x is the number reported in Table 5 scaled by 100). Solving for the Ask/Bid ratio gives $(\text{Ask/Bid} = (1+2x)/(1-2x) \approx 4x$, which follows from taking a first order Taylor approximation for the log of the Ask/Bid ratio around $x = 0$. With a median spread of 0.2, the Ask/Bid ratio is approximately $4 \times 0.2 = 0.8$, or 80 basis points.

We conclude the descriptive statistics with Table 6. It reports key statistics for the target and the mimicking portfolio. The number of unique targets increases from 3 in 2010 to 16 in 2020. In light of the number of possible targets, which increases from 20 to 71, the result is surprising: a few fixed income ETFs are repeatedly priced differently than their NAV. The fraction of mispriced ETFs increases slightly over the same period from 15 percent (3/20) to 23 percent (16/71). Hence, one out of five or six ETFs in the fixed income market is at times sufficiently mispriced for us to exploit it. This suggests that current arbitrage strategies are insufficient to ensure efficiency in the fixed income ETF market.

[Insert Table 6 here]

The last three columns of Table 6 shows the asset class composition of the mimicking portfolio. Roughly 20 to 30 percent is fixed income, 10 to 15 percent is government bonds and money market funds, and the remainder is equity. The combination of asset classes is the result of the genetic algorithm and natural selection. More price variation within the mimicking portfolio reduces collinearity and yields a better fit.

3.2 The Statistics of the Arbitrage Portfolio

3.2.1 Descriptive statistics

Table 7 reports summary statistics for the arbitrage portfolio by year from 2010 to 2020. The annualized average daily returns are between 1 percent (2010) and 56.9 percent (2020). The

corresponding standard deviations are between 1.6 percent (2017) and 11.8 percent (2010). From 2013, the mean return is at least four times greater than the standard deviation.

[Insert Table 7 here]

Figure 3 compares the value of one dollar invested in the arbitrage portfolio in 2010 with the same equity investment. As expected from the statistics on the first and second moment of the returns, it shows that the arbitrage portfolio is effectively a bank account with a twist; its interest rate is comparable to the average return on equities. The annualized Sharpe ratios are astronomical and range from 0.1 (2010) to 8.9 (2015) with a median of 5.3.

[Insert Figure 3 here]

The other statistics in Table 7 are: i) the first-order autocorrelations (AC1). These estimates are negative or close to zero. In years with non-zero autocorrelation, returns above the mean in one day are, on average, followed by returns below the mean in the subsequent day. The most negative first-order autocorrelation estimates are in 2012, 2013, and 2020, ranging from -0.3 to -0.2. ii) The spreads (i.e., $\ln(\text{Price}_{i,t}^{ETF}/\text{NAV}_{i,t})$) are between -0.3 and -0.9 with the median of -0.45. Thus, a typical target trades at a discount of 45 basis points relative to its NAV. The reason they are all negative is that we always sell the target short and therefore do not consider the cases when the ETF is cheap relative to its NAV. iii) The R^2 of the NNLS regression of the ETFs in the mimicking portfolio is between 18.4 (2014) and 97.7 (2011).

3.2.2 Transaction cost

Although our method circumvents illiquid securities, it is only an alternative to traditional arbitrage if the gains from using it exceed its costs. For reasons we now explain, the returns remain high after transaction costs, which we expect not to exceed 4 percent (annualized). The seemingly low transaction costs are a byproduct of the arbitrage strategy and the ETF

market. NNLS only allows for long positions in the mimicking portfolio, which can be traded with zero commissions through most brokers. The transaction costs, therefore, come from the short position that is financed at fluctuating rates. In calm markets, lending rates for high yield fixed-income ETFs can fluctuate between 0.25 and 1 percentage point. During the COVID crisis and subsequent market distress in early 2020, lending rates increased to about 4 percentage points. Thus, our estimate of financing the short position at 4 percent at all times is conservative. With a 4 percent cost estimate and a portfolio return of r_a, transaction costs reduce the Sharpe ratios by $-0.04/r_a$ percent.[5] For instance, in 2013, 2018, and 2020, the Sharpe ratios of the arbitrage strategy were at the median of 5.2. After transaction cost they would still be 3.1, 4.2, and 4.8.[6]

3.3 Risk-Adjusted Returns

We proceed by adjusting the return on the arbitrage portfolio for equity and bond factors. The factor portfolios are from Turan Bali's and Kenneth French's websites.[7] Because the bond factors are only available at the monthly frequency, we convert our daily returns into monthly returns when we use those factors. Table 8 presents the results.

[Insert Table 8 here]

The main message is that the return on the arbitrage portfolio is market neutral. After adjusting for bond and equity factors the monthly return on the arbitrage portfolio is approximately 1.35 percentage points. Annualized, this is 16.2 percentage points. Standard errors are low, varying from 0.148 to 0.166, depending on the factors. The resulting t-statistics are between 8 and 9. The portfolio's loadings on all factors are small in magnitude and statistically insignificant. Adjusted R^2 are negative in all specifications.

[5] This follows from taking the derivative of the Sharpe Ratio with respect to realized return.

[6] We obtain these by adjusting the Sharpe ratios with $-0.04/r_a$, taking r_a from Table 7.

[7] The bond and equity factors are from Turan Bali's and Kenneth French's websites. Bonds: https://sites.google.com/a/georgetown.edu/turan-bali/data-working-papers?authuser=0. Equity: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 9 reports the results of regressing the return on the arbitrage portfolio on three equity factor models at the daily frequency. The results are reminiscent of those obtained at the monthly frequency. After adjusting for equity factors, the daily return on the arbitrage portfolio is about 7.1 basis points. Annualized, this is 17.9 (252×7.1) percentage points. The t-statistics of the regression intercepts remain high at 7.65 or greater. Adjusted R^2 are now positive but less than 0.02 in all specifications.

[Insert Table 9 here]

3.4 Understanding Arbitrage Profit

The profits from arbitrage trading reveal an inefficiency in the fixed income ETF market. The risk-adjusted returns we document are exceptional and cannot be explained by transaction costs or common risk factors. In this section, we use multiple specifications to shed light on the dynamics of the arbitrage profits.

3.4.1 Time-series variation in arbitrage profits

We start by regressing the contemporaneous return on the arbitrage portfolio proxies for the availability of arbitrage capital. As proxies, we include changes in the VIX (ΔVIX) and the TED (ΔTED) spread. The VIX index is one common proxy for the availability of arbitrage capital (Petajisto, 2017). Another is the Ted spread, that is, the difference between the three-month LIBOR and T-bill rates. It is an estimate of the interest rate financial institutions pay over the risk-free rate for unsecured lending (see e.g., Brunnermeier et al., 2008). Both variables are normalized to have a mean of zero and a standard deviation of one. We also include the return on the U.S. equity market as independent variable. It is meant to capture the state of the overall economy.

Table 10 presents the results. The regression coefficients for ΔVIX and ΔTED are about 2 basis points with standard errors of almost the same magnitude. In comparison, the

mean daily return of the arbitrage portfolio is 7 basis points. Thus, shocks to the available arbitrage capital, measured as realizations two standard deviations away from the mean, are associated with arbitrage profits about 50 percent above its mean.

[Insert Table 10 here]

3.4.2 Is the return on the arbitrage portfolio predictable?

The second column of Table 10 presents the results from regressing the return on the arbitrage portfolio tomorrow on two independent variables and their interaction. The independent variables are: R^2 measuring how well the mimicking portfolio replicates the short position, the ETF premium, and their interaction. The coefficient on R^2 is negative and statistically significant. Therefore, the greatest gains come from short positions that are difficult to replicate. A worse fit of the mimicking portfolio means more risk for the arbitrageurs. Furthermore, the coefficient of the ETF premium of the target is small and insignificant. Hence, the level of mispricing today is not related to the profit realized the next day. The last row shows the result for the interaction term. Because R^2 is nonnegative and the ETF premium is always negative, the product of the two is always negative. The positive coefficient suggests that a combination of a high R^2 and a large ETF premium gives little profit. This finding may be interpreted as deals that are too good to be true are not profitable.

3.4.3 The dynamics of the ETF premium

In the third and fourth columns of Table 10, we study the dynamics of the ETF premium. The largest mispricing, measured by the ETF premium, occurs at times when the equity market drops in value. This is intuitive. In times of distress, when volatility rises, all markets tend to get correlated, and this will impact fixed income markets. Since the ETF is traded on an exchange and is more liquid than the underlying bonds, it will absorb new information faster, and this will increase the premium as it takes longer for the underlying bonds to correct. As time passes, the premium reverts to normal levels. The second and third rows of

the third column show that there is no relationship between the available arbitrage capital and the ETF premium. In the fourth column, we predict the ETF premium. The R^2 from this regression is 0.37 and the coefficient of the lagged ETF premium is 0.62 and highly statistically significant. The ETF premium is predictable and persistent.

3.4.4 The dynamics of R^2

In the last two columns, we run the same regressions as for the ETF premium but with R^2 as the dependent variable. A notable difference between the ETF premium and R^2 is that R^2 forecasts trading profits (the second column of Table 10 shows this). Thus, the lower R^2 today, the higher is the expected return from the arbitrage trade. The fifth column shows that R^2 is on average lower when the equity market decreases in value than when it increases. Taken together, arbitrage trades based on mispriced ETFs that are difficult to replicate on days the equity market drops in value give the highest realized return. The last column shows that R^2 tends to be low after the equity market drops or the Ted spread increases.

3.4.5 Does it matter how liquid the target is?

We regress the return on the arbitrage strategy on two measures of how liquid the target is: the log of traded volume and the normalized bid-ask spread. Both measures are normalized to have a mean of zero and a standard deviation of one. Table 11 presents the results.

[Insert Table 11 here]

In the first three columns, we study contemporaneous relationships. The coefficient of bid-ask spread is 0.042 and statistically significant. It shows that the daily return on the arbitrage portfolio is 4.2 basis points higher when we short a target that has a bid-ask spread one standard deviation above the mean spread. Given that the unconditional mean return is 7.7 basis points, the coefficient is large. In the last three columns, we study how illiquidity relates to the expected return from the arbitrage trade. Both coefficients are statistically

significant and tell the same story: entering an arbitrage trade with a less liquid target increases the expected return on the trade.

Taken together, our results indicate that arbitrage profits are higher in periods in which it is more difficult for arbitrageurs to exploit mispricing and in more risky arbitrage trades. Difficult times are periods in which available arbitrage capital is scarce. Risky trades are positions with a more noisy mimicking portfolio or a less liquid target.

3.5 Alternative Methods

Identifying a replicating portfolio from a large set of traded assets by combining natural selection (a genetic algorithm) with NNLS gives three main benefits. The first is stable portfolio weights that are all positive. This is a direct consequence of using NNLS instead of OLS. The second benefit comes from diversification. The last benefit is the genetic algorithm's ability to find "robust" relationships in the data. We expect this feature to prevent losses during financial market turmoil when historical correlations fail.

We illustrate the two last benefits by comparing the mean return, standard deviation, and annualized Sharpe ratios of the main specification with two alternatives. The first alternative ("1 ETF") sells the same target but takes a long position in the ETF with the highest correlation with the target. The second alternative ("8 ETF") uses a mimicking portfolio based on the eight ETFs with the highest correlation with the target. Table 12 presents the results.

[Insert Table 12 here]

Diversification is essential for arbitrage profits. The Sharpe ratio of the alternative ("1 ETF") is always considerably below the main specification. During periods with large positive shocks to the Ted spread, the alternative ("1 ETF") loses money. Second, the genetic algorithm protects the profits from crashes during financial market turmoil. Unlike the second alternative ("8 ETF"), which is also well diversified, the Sharpe ratio of the main

specification is similar in good and bad times.

4 Conclusion

We develop a methodology that arbitrageurs can use to exploit mispriced contingent claims. Unlike traditional arbitrage, our method does not involve taking a direct position in the underlying assets. Rather, we combine natural selection (a genetic algorithm) with NNLS to identify a mimicking portfolio of liquid assets that work as a substitute for underlying assets. As a result, our method makes it possible to trade on market inefficiencies while circumventing the *de facto* non-tradability of the underlying securities.

We apply our method to the US market for fixed income ETFs. Our annual average daily returns are between 1 percent (2010) and 56.9 percent (2020). The corresponding standard deviations are between 1.6 percent (2017) and 11.8 percent (2010). From 2013, the mean return is at least four times greater than the standard deviation. The annualized Sharpe ratios are between 0.1 (2010) and 8.9 (2015) with a median of 5.3. After the transaction cost, we estimate that the Sharpe ratio of our arbitrage strategy is at least 3.

While these profits are attractive based on the estimated Sharpe ratios, a large part of the profits are made during periods in which arbitrage capital is expensive and on more risky trades. Risky trades mean holding a more noisy mimicking portfolio or a less liquid target.

Our results have implications for the arbitrage theory of ETFs. We show that, without trading the underlying, arbitrageurs can profit from the difference between the ETF price and the NAV. Hence, the important role of authorized participants reduces as we have more and more liquid ETFs since one investor can arbitrage away deviations from the NAV by using the other ETFs. Therefore, when considering how ETF issuances affect their underlying securities, we must consider aggregate deviations.

The methodology we develop can be applied in many other settings. Detzel et al. (2021) show that many anomalies discovered in the equity market (e.g., momentum) are unlikely

profitable after transaction costs. An interesting avenue for future research is whether our methodology can be applied to replicate equity anomalies with liquid ETFs.

5 Appendix: Hyperparameters

The genetic algorithm relies on several hyperparameters. Here we list the hyperparameters and explain the reasoning behind each parameter choice. The hyperparameters are summarized in Table 1.

Table 1 Hyperparameters in the genetic algorithm

ETFs	population size	Crossover	Mutation prob.	No. of gen.
8	100	Yes	0.083	500

The number of subsets of ETFs in each generation. The set of subsets of ETFs constitutes a gene pool. As for gene pools in general, this set needs sufficient genetic variation to facilitate natural (fitness) selection. A large population tends to have more genetic variation than a smaller one. From a natural selection perspective, smaller populations evolve faster as attractive genes dominate faster. From a computational perspective, run times increase considerably as the number of subsets of ETFs grows. While the number of subsets needed to ensure sufficient genetic variation may vary from domain to domain, the literature (see e.g., Marsland, 2009) indicate that around 100 is often sufficient. Preruns are the litmus test regarding the gene pool size. In this case, we did preruns against a random ETF and a population size of a 100 (with the mutation rate given below) and found rapid increases in fitness from generation to generation. Of course, with such preruns, we can only observe the level and consistency in the improvement of fitness (R^2), not the maximum fitness level (here the highest R^2 of a mimicking portfolio)

The mutation rate. The mutation rate, which is the probability that any given ETF in the offspring is replaced by a random draw of an ETF from the set of all possible ETFs is $p = 1 - \frac{1}{\sqrt[8]{2}} \approx 0.083$. This probability is chosen to obtain a probability of 50 percent that

the offspring has at least one mutation. A considerable mutation rate, as here, is important in the case of a small population like the one we have. It ensures that essential genes, not present or lost in the population, could be (re)introduced. As with the number of subsets in the population, we have some discretion concerning the actual values. Again, we rely on preruns against a random target ETF. These preruns showed rapid increases in fitness from generation to generation with this mutation rate (and a population of 100 ETF subsets.)

The crossover rule. Our crossover rule, which is the recombination of two subsets of ETFs, relies on a threeway division of the 100 subsets. First, we select the 33 most fit individuals and pair those with the group of the next 33, that is, the 34 to 66 most fit individuals. We form 33 pairs by pairing after fitness. In other words, we pair the most fit in the first group with the fittest in the second group. Each pair will serve as parents for an offspring. A random draw of 4 ETFs creates the offspring from parent 1, and the remaining 4 are drawn from parent 2 ETFs not previously drawn from parent 1. This ensures that the new individual has 8 distinct ETFs. Although the details of the recombination, such as the number of pairs, can vary, our choice of pairing is standard (see Sommervoll and Sommervoll, 2019). The key idea is to replace many subsets each generation without losing too much genetic variation. As genetic variation also depends on the mutation rate, these two hyperparameters must be seen in combination. Preruns with a random target ETF showed that this recombination rule ensured rapid generational gain in fitness.

The number of generations. This hyperparameter is set to 500. This number is excessive because preruns of a mimicking portfolio for a given random ETF tend to have modest to non-existent improvements for the 100-200 last runs. As for the other hyperparameters, we have some leeway regarding the stopping criterion; and modest to large deviations from 500 have limited impact on the results.

The number of ETFs in the mimicking portfolio. The number of ETFs required for an adequate mimicking portfolio is a hyperparameter that requires tuning as it needs to be apriori clear what is a sufficient number of ETFs to provide a suitable mimicking portfolio. Overall, the more regressors, the better the in-sample fit. At the same time, more regressors translate into more positions and, consequently, more trading. We try mimicking portfolios of 3,4,5,6,7,8,9,10 ETFs for a random target ETF using an earlier part of the sample to prevent data leakage. It was clear that, at best, modest improvements in fitness were realized going beyond 8 ETFs. Hence, we chose up to 8 (in many cases the NNLS selects fewer ETFs than 8) ETFs in our mimicking portfolio.

References

Bai, Jennie, Turan G. Bali, and Quan Wen, "Common risk factors in the cross-section of corporate bond returns," *Journal of Financial Economics*, 2019, *131* (3), 619–642.

Ban, Gah-Yi, Noureddine El Karoui, and Andrew EB Lim, "Machine learning and portfolio optimization," *Management Science*, 2018, *64* (3), 1136–1154.

Ben-David, Itzhak, Francesco Franzoni, and Rabih Moussawi, "Do ETFs Increase Volatility?," *The Journal of Finance*, 2018, *73* (6), 2471–2535.

Bessembinder, Hendrik and William Maxwell, "Markets: Transparency and the corporate bond market," *Journal of economic perspectives*, 2008, *22* (2), 217–234.

Brogaard, Jonathan and Abalfazl Zareei, "Machine learning and the stock market," 2021.

Brunnermeier, Markus, Stefan Nagel, and Lasse Pedersen, "Carry Trades and Currency Crashes," NBER Working Papers 14473, National Bureau of Economic Research, Inc 2008.

Dannhauser, Caitlin D and Saeid Hoseinzade, "The Unintended Consequences of Corporate Bond ETFs: Evidence from the Taper Tantrum," *The Review of Financial Studies*, 03 2021, *35* (1), 51–90.

Detzel, Andrew L., Robert Novy-Marx, and Mihail Velikov, "Model Selection with Transaction Costs," 2021. Working paper.

Easley, David, Marcos López de Prado, Maureen O'Hara, and Zhibai Zhang, "Microstructure in the machine age," *The Review of Financial Studies*, 2021, *34* (7), 3316–3363.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, "Taming the factor zoo: A test of new factors," *The Journal of Finance*, 2020, *75* (3), 1327–1370.

Friedman, Milton, *Essays in Positive Economics*, University of Chicago Press, 1953.

Harrison, J. Michael and David M. Kreps, "Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations," *The Quarterly Journal of Economics*, 1978, *92* (2), 323–336.

Laipply, Stephen and Ananth Madhavan, "Pricing and Liquidity of Fixed Income ETFs in the Covid-19 Virus Crisis of 2020," *Journal of Beta Investment Strategies*, 2020, *11* (3), 7–19.

Madhavan, Ananth and Aleksander Sobczyk, "Historical returns of the market portfolio," *Journal of Investment Management*, 2016, *14* (2), 86–102.

Marsland, Stephen, *Machine Learning: An Algorithmic Perspective*, 1st ed., Chapman & Hall/CRC, 2009.

Newey, Whitney K. and Kenneth D. West, "Automatic Lag Selection in Covariance Matrix Estimation," *The Review of Economic Studies*, 1994, *61* (4), 631–653.

Petajisto, Antti, "Inefficiencies in the pricing of exchange-traded funds," *Financial Analysts Journal*, 2017, *73* (1), 24–54.

Roll, Richard and Akshay Srivastava, "Mimicking portfolios," *The Journal of Portfolio Management*, 2018, *44* (5), 21–35.

Shim, John J. and Karamfil Todorov, "ETFs, Illiquid Assets, and Fire Sales," *Working Paper*, 2022.

Sommervoll, Åvald and Dag Einar Sommervoll, "Learning from man or machine: Spatial fixed effects in urban econometrics," *Regional Science and Urban Economics*, 2019, *77*, 239–252.

Todorov, Karamfil, "The anatomy of bond ETF arbitrage," Technical Report 1 2021.

Tucker, Matthew and Stephen Laipply, "Bond Market Price Discovery: Clarity Through the Lens of an Exchange," *The Journal of Portfolio Management*, 2013, *39* (2), 49–62.

Tables and Figures

Tables

Table 2 Specification of the genetic algorithm

population size (N)	Crossover	Mutation prob.	No. of generations
100	Yes	0.083	500

Table 3
ETFs by Lipper Class

This table reports the Lipper Class definitions of the ETFs in the sample. We group the Lipper Class definitions into three asset classes, which are corporate credit (C), government and money market funds (G), and domestic (U.S.) equity (E). Nobs refers to the number of unique securities within each asset class. for the period

	Lipper Class		
Unique ETFs	Corporate Credit (C)	Government & Money market (G)	Domestic Equity (E)
2002-2020	94	78	823
2010-2020	94	78	778
	A	IUT	AU
	BBB	SUS	H
	CPB	SUT	FS
		IUG	NR
		SIU	RE
		GUS	TK
		GUT	UT
		CAM	CG
		CTM	CMD
		MAM	CS
		MIM	ID
		NUM	BM
		NYM	TL
		OHM	SP
		OTM	MC
		PAM	SG
		IMM	MR
		ITE	CA
		ITM	G
		IUS	GI
		MM	LSE
		TEM	EMN
		USS	ABR
		UST	DL
			DSB
			EI

Table 4
ETFs by Lipper Class

This table reports the unique ETFs by asset classes and calendar year.

	Lipper Class		
Unique ETFs	Corporate Credit	Government & Money market	Domestic Equity
2002	1	3	51
2003	1	4	51
2004	1	4	60
2005	1	4	65
2006	1	4	118
2007	6	11	214
2008	7	15	248
2009	12	28	243
2010	20	35	264
2011	21	40	292
2012	28	44	291
2013	32	46	292
2014	37	49	309
2015	45	49	349
2016	44	49	402
2017	50	49	410
2018	65	51	489
2019	65	52	500
2020	71	66	538

Table 5
ETFs by Lipper Class

This table reports the distribution of volume and bid-ask relative to close price by asset classes and calendar year. We report the statistics at the asset class level, which are corporate credit (C), government and money market funds (G), and domestic (U.S.) equity (E).

Daily Volume (thousands)

	10th Percentile			Median			90th Percentile		
	C	G	E	C	G	E	C	G	E
2010	0.6	1.0	1.1	17.4	18.0	34.0	319	638	2784
2011	0.7	0.8	0.5	20.0	20.4	24.8	265	616	1711
2012	0.5	0.3	0.1	30.2	18.1	14.4	428	353	1044
2013	0.2	0.3	0.2	41.0	21.0	20.7	482	423	1086
2014	0.5	0.2	0.4	23.6	17.1	23.1	326	317	1119
2015	0.0	0.2	0.3	17.6	27.1	26.7	334	423	1231
2016	0.6	0.3	0.0	24.4	38.2	19.3	410	638	1190
2017	0.6	0.8	0.2	27.7	34.2	20.3	330	677	853
2018	0.0	1.7	0.5	21.4	71.4	26.8	392	1159	1292
2019	0.2	1.5	0.5	22.9	89.5	19.1	621	1711	1072
2020	1.2	0.6	1.0	36.4	80.6	36.7	860	2148	2652

(Ask - Bid)/ Close Price × 100

	10th Percentile			Median			90th Percentile		
	C	G	E	C	G	E	C	G	E
2010	0.09	0.06	0.46	0.39	0.34	1.41	0.96	1.46	3.65
2011	0.17	0.06	0.37	0.43	0.36	1.38	1.26	1.61	3.96
2012	0.11	0.04	0.21	0.29	0.27	0.99	0.77	1.09	2.66
2013	0.09	0.04	0.27	0.31	0.26	0.91	1.01	1.17	2.41
2014	0.10	0.05	0.28	0.29	0.24	0.92	0.77	1.05	2.58
2015	0.10	0.06	0.28	0.37	0.33	1.09	0.84	1.39	3.02
2016	0.13	0.06	0.17	0.37	0.29	1.02	0.81	1.20	3.30
2017	0.09	0.04	0.16	0.26	0.21	0.76	0.57	0.84	2.17
2018	0.08	0.04	0.22	0.25	0.18	1.14	0.55	0.74	3.47
2019	0.08	0.03	0.23	0.23	0.18	0.95	0.50	0.81	2.60
2020	0.11	0.02	0.44	0.33	0.19	1.73	1.17	1.19	5.59

Table 6
Portfolio Composition

This table reports statistics for the target and the composition of the replicating portfolio by year. Prem is defined as the differences in logs between the ETF price (PRC) and its corresponding net asset value (NAV) multiplied by 100. The column labeled as Mean refers to mean daily time-series average within the year. SD refers to the corresponding standard deviaiton. Mimicking Portfolio presents summary statistics of the fraction of the mimicking porfolio in each of the three asset classes.

	Target Statistics			Mimicking Portfolio		
	Unique	Prem		Corporate	Government &	Domestic
year	Targets	Mean	SD	Credit	Money market	Equity
2010	3	-0.31	0.41	0.23	0.11	0.66
2011	8	-0.54	0.41	0.27	0.17	0.56
2012	6	-0.91	0.36	0.32	0.16	0.53
2013	12	-0.45	0.15	0.29	0.08	0.63
2014	12	-0.44	0.20	0.33	0.13	0.54
2015	14	-0.42	0.20	0.37	0.18	0.45
2016	13	-0.58	0.30	0.35	0.13	0.52
2017	11	-0.38	0.10	0.48	0.13	0.39
2018	21	-0.31	0.24	0.41	0.09	0.50
2019	15	-0.46	0.20	0.35	0.08	0.57
2020	16	-0.90	1.13	0.34	0.07	0.60

Table 7
Descriptive Statistics

This table reports statistics for the arbitrage portfolio by year. The column labeled as Mean refers to the average daily return multiplied 252 (trading days) and is reported in percentage points. SD refers to the corresponding standard deviaiton. AC1 is the first order autocorrelation in daily returns multiplied by 100. Spread denotes is defined as the differences in logs between the ETF price (PRC) and its corresponding net asset value (NAV) multiplied by 100. $R2$ is the average $R2$ in a regressing with the target's ETF price as the dependent variable and the ETFs in the replicating portfolio as the independent variables using 500 trading days. SR refers to the Sharpe Ratio and is calcualted as the mean daily return scaled by the daily standard deviation multiplied by the square root of 252 (trading days).

Year	Mean	SD	AC1	Premium	$R2$	SR
2010	1.0	11.8	0.2	-0.31	95.8	0.1
2011	12.4	3.5	4.5	-0.54	97.7	3.6
2012	8.7	3.9	-22.3	-0.91	97.5	2.3
2013	13.9	2.7	-28.8	-0.45	79.0	5.2
2014	18.3	2.1	-2.2	-0.44	18.4	8.8
2015	20.8	2.3	-4.6	-0.42	82.0	8.9
2016	28.2	3.7	-3.0	-0.58	81.7	7.6
2017	14.0	1.6	-16.7	-0.38	62.0	8.8
2018	25.0	4.6	-15.6	-0.31	93.4	5.2
2019	15.0	2.4	-11.8	-0.46	96.4	6.3
2020	56.9	10.8	-29.1	-0.90	95.2	5.3

Table 8
Monthly Bond and Equity Factor Regressions

The table reports OLS regressions of monthly returns (in percentages points) on the arbitrage portfolio on different bond and equity factor models for the period January 2010 to December 2019. We report the intercept (α), the factor loadings, and the respective OLS standard errors. Equity factors are downloaded from Kenneth French's website. The CAPM refers to the equity market factor, FF3 = $\{MKT, SMB, HLM\}$,and FF5 = $\{MKT, SMB, HLM, CMA, RMW\}$. Bond factors are downloaded from Turan Bali's website. MKTBond is the bond market factor, DRF refers to the downside risk factor, LRF is the liquidity risk factor, and CRF is the the final credit risk factor. We refer to the original paper for additional details on the bond factors (Bai et al., 2019). The superscript * denotes statistical signiifcance with *p<0.1; **p<0.05; ***p<0.01, respectively.

Intercept (α)	1.336***	1.344***	1.366***
	(0.148)	(0.157)	(0.166)
Bond factors:			
Bond Market	−0.050	−0.013	0.026
	(0.141)	(0.189)	(0.200)
DRF		−0.002	−0.038
		(0.111)	(0.115)
CRF		0.031	0.044
		(0.109)	(0.112)
LRF		−0.085	−0.095
		(0.221)	(0.232)
Equity factors:			
Equity Market			−0.0002
			(0.043)
SMB			0.074
			(0.076)
HML			0.112
			(0.083)
CMA			−0.034
			(0.133)
RMW			0.132
			(0.111)
Observations	119	119	119
Adjusted R^2	−0.007	−0.032	−0.034

Table 9
Daily Equity Factor Regressions

The table reports OLS regressions of the daily returns (in percentages points) on the arbitrage portfolio on different equity factor models for the period January 2010 to June 2020. We report the intercept (α) the respective Newey and West (1994) standard errors. All factor models are downloaded from Kenneth French's website The CAPM refers to the market factor, FF3 = $\{MKT, SMB, HLM\}$,and FF5 = $\{MKT, SMB, HLM, CMA, RMW\}$. The superscript * denotes statistical signiifcance with *p<0.1; **p<0.05; ***p<0.01, respectively.

Intercept (α)	0.07137***	0.07176***	0.07194***
	(0.00928)	(0.00761)	(0.00820)
Benchmark model	CAPM	FF3	FF5
Control variables: Lagged factor returns	Yes	Yes	Yes
Observations	2,639	2,639	2,639
Adjusted R^2	0.01147	0.01068	0.01660

Table 10
The Drivers of Arbitrage Profits

The table reports OLS regressions of the daily observations for a set of specifications for the period January 2010 to June 2020. r_t refers to the return on the arbitrage portfolio at time t. Prem denotes is defined as the differences in logs between the ETF price (PRC) and its corresponding net asset value (NAV) multiplied by 100. $R2$ is the average $R2$ in a regressing with the target's ETF price as the dependent variable and the ETFs in the replicating portfolio as the independent variables using 500 trading days. We report the intercept and the slope coefficients and their respective Newey and West (1994) standard errors. Equity Market$_t$ refers to the equity market factor. ΔVIX refers to the first difference of the log of the CBOE Volatility Index (VIXCLS available from), ΔTED is defined as the first difference of the log of the spread between 3-Month LIBOR based on US dollars and 3-Month Treasury Bill. Both ΔVIX and ΔTED are noramlized to have a mean of zero and a standard deviation of one. The superscript * denotes statistical signiifcance with *p<0.1; **p<0.05; ***p<0.01, respectively.

Outcome:	r_t	r_{t+1}	$Prem_t$	$Prem_{t+1}$	$R2_t$	$R2_{t+1}$
Equity Market$_t$	0.029		-0.276^{***}	0.109	0.046^{**}	0.034^{*}
	(0.039)		(0.102)	(0.110)	(0.022)	(0.017)
ΔVIX_t	0.019		-0.090	0.094	0.052	0.048
	(0.022)		(0.077)	(0.080)	(0.034)	(0.032)
ΔTED_t	0.018		0.002	0.004	-0.017	-0.022^{**}
	(0.013)		(0.026)	(0.012)	(0.011)	(0.011)
$R2_t$		-0.034^{**}				0.181^{**}
		(0.014)				(0.083)
$Prem_t$		-0.003		0.620^{***}		
		(0.009)		(0.058)		
$R2_t \times Prem_t$		0.059^{***}				
		(0.021)				
Intercept	0.070^{***}	0.075^{***}	-0.000	-0.000	-0.000	-0.000
	(0.007)	(0.009)	(0.040)	(0.019)	(0.020)	(0.019)
Observations	2,573	2,572	2,573	2,572	2,572	2,572
Adjusted R^2	0.004	0.001	0.046	0.374	0.0003	0.033

Table 11
Target Liqudity and Arbitrage Profits

The table reports OLS regressions of the daily returns from the arbitrage portfolio on log volume and the bid-ask spread for the period January 2010 to June 2020. r_t refers to the return on the arbitrage portfolio at time t. We report the intercept and the slope coefficients and their respective Newey and West (1994) standard errors. The superscript * denotes statistical signiifcance with *p<0.1; **p<0.05; ***p<0.01, respectively.

Outcome:	r_t	r_t	r_t	r_{t+1}	r_{t+1}	r_{t+1}
Intercept	0.077***	0.077***	0.077***	0.077***	0.077***	0.077***
	(0.008)	(0.007)	(0.007)	(0.008)	(0.006)	(0.007)
$\ln(Volume)$	−0.006		−0.009	−0.017***		−0.020***
	(0.008)		(0.008)	(0.006)		(0.006)
$\frac{\text{Ask High}_t - \text{Bid Low}_t}{\text{Price}_t}$		0.042***	0.042***		0.052***	0.053***
		(0.010)	(0.011)		(0.019)	(0.018)
Observations	2,999	2,999	2,999	2,998	2,998	2,998
Adjusted R^2	0.00003	0.016	0.016	0.002	0.024	0.027

Table 12
Sharpe Ratios for Alternative Methods

This table reports the daily mean return and its standard deviation in percentage points together with the annualized Sharpe Ratios (i.e., the ratio of mean to standard deviation multiplied by the square root of 252 (trading days). GA refers to the baseline strategy. 1 ETF include only the ETF with the highest correlation with the short position in the mimicking portfolio. 8 ETF include 8 of the ETFs with the highest correlation. The short position is the same for all specifications. Column 4,5,6 report the corresponding statistics for the U.S. equity market, and percentage change in the VIX and the Ted spread. The first panel report the resulst for the full sample. The second and third panel report the results for good and bad times based on percentage change in the VIX. The fourt and fifth panel report the correpsponding results but using the percentage change in the Ted spread. Bad times are defined as periods with the 5 percent largest increase in the VIX and the Ted spread. Good times is the complement.

Jan 2010 - Jun 2020

	GA	1 ETF	8 ETF	MKT	dVIX	dTED
Mean	0.070	0.051	0.062	0.047	-0.008	-0.008
SD	0.335	1.413	0.665	1.128	7.946	6.995
SR	3.323	0.578	1.477	0.663		

Returns in bad times (VIX)

	GA	1 ETF	8 ETF	MKT	dVIX	dTED
Mean	0.076	0.235	0.001	-2.208	21.113	1.011
SD	0.253	1.882	1.157	1.682	8.895	9.807
SR	4.796	1.984	0.012	-20.833		

Returns in good times (VIX)

	GA	1 ETF	8 ETF	MKT	dVIX	dTED
Mean	0.070	0.042	0.065	0.166	-1.123	-0.061
SD	0.338	1.384	0.629	0.953	6.127	6.813
SR	3.270	0.479	1.643	2.765		

Returns in bad times (TED)

	GA	1 ETF	8 ETF	MKT	dVIX	dTED
Mean	0.141	-0.202	0.039	-0.259	1.289	16.732
SD	0.619	1.944	1.129	2.135	10.995	9.276
SR	3.631	-1.652	0.554	-1.927		

Returns in good times (TED)

	GA	1 ETF	8 ETF	MKT	dVIX	dTED
Mean	0.066	0.064	0.063	0.063	-0.074	-0.855
SD	0.313	1.380	0.633	1.050	7.757	5.673
SR	3.366	0.740	1.580	0.946		

Figures

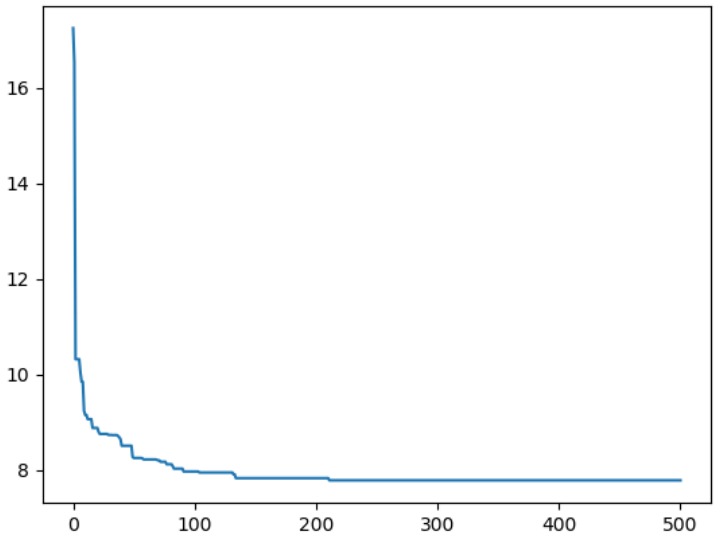

Figure 2 Mean absolute basispoint deviation best mimicking portfolio by generation

This figure illustrates the genetic algorithm. Its purpose is to find the combination of 8 ETFs that gives the best "fit". Fitness is measured by the Mean absolute deviation (MAD) between the price of the target and the price of the mimicking portfolio. The MAD is based on prices one day after portfolio formation and reported in basis points. The y-axis shows the MAD and the x-axis shows the number of generations.

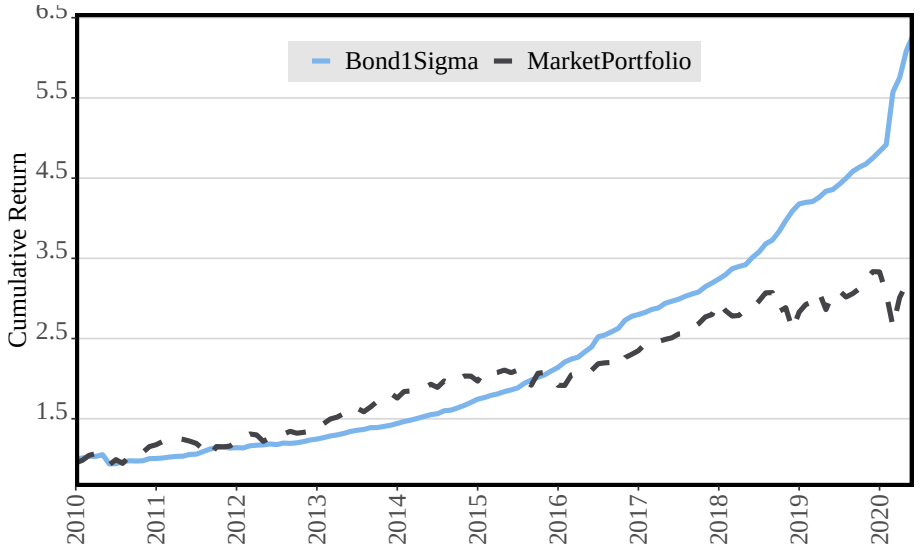

Figure 3 Cumulative Performance

This figure plots the value of one dollar invested in January 2010 in two portfolios. The first portfolio ("Bond1Sigma") selects a target to sell short among ETFs whose premium (i.e., $\ln(PRC/NAV)$) is one standard deviation above the cross-sectional mean in a given day. The second portfolio is the U.S. Equity market.