

Statistical Arbitrage Risk Premium by Machine Learning

Raymond C. W. Leung and Yu-Man Tam ^{*}

March 2021

Abstract

How to hedge factor risks without knowing the identities of the factors? We first prove a general theoretical result: even if the exact set of factors cannot be identified, any risky asset can use some portfolio of similar peer assets to hedge against its own factor exposures. A long position of a risky asset and a short position of a “replicate portfolio” of its peers represent that asset’s factor residual risk. We coin the expected return of an asset’s factor residual risk as its *Statistical Arbitrage Risk Premium* (SARP). The challenge in empirically estimating SARP is finding the peers for each asset and constructing the replicate portfolios. We use the elastic-net, a machine learning method, to project each stock’s past returns onto that of every other stock. The resulting high-dimensional but sparse projection vector serves as investment weights in constructing the stocks’ replicate portfolios. We say a stock has high (low) *Statistical Arbitrage Risk* (SAR) if it has low (high) R-squared with its peers. The key finding is that “unique” stocks have both a higher SARP and higher excess returns than “ubiquitous” stocks: in the cross-section, high SAR stocks have a monthly SARP (monthly excess returns) that is 1.101% (0.710%) greater than low SAR stocks. The average SAR across all stocks is countercyclical. Our results are robust to controlling for various known priced factors and characteristics.

Keywords: cross-sectional, elastic-net, empirical asset pricing, machine learning, portfolio construction, statistical arbitrage risk premium, statistical arbitrage risk

JEL classification: G11, G12

^{*}Leung is with the Cheung Kong Graduate School of Business, and Tam is with the Office of the Comptroller of the Currency, U.S. Department of Treasury. This paper is a substantially revised version of a paper previously circulated under the title “Asset Insurance Premium in the Cross-Section of Asset Synchronicity”. Please send all correspondences to Leung at raymond.chi.wai.leung@gmail.com. Part of this research is done while Leung was visiting UC Berkeley. We thank Robert M. Anderson, Lisa Goldberg, Alex Michaelides, Neal Stoughton, Baolian Wang, Hongjun Yan, and Wenhao Yang for helpful discussions. We also thank the seminar participants of the Five-Star Forum, Internal Machine Learning Workshop at the OCC, SWUFE-CDAR 2018 Symposium and the 2019 FMA Annual Meeting in New Orleans for helpful comments and suggestions. We thank an anonymous referee who made several suggestions that significantly improved the paper. Views and opinions expressed are those of the authors and do not necessarily represent official positions or policy of the OCC. All errors are ours.

Given any stock, how can one hedge against its factor risks? This question is simple to answer with a linear factor model structure. For instance, under the Markowitz mean-variance portfolio theory and its resulting equilibrium capital asset pricing model (CAPM) (Sharpe (1964), Lintner (1965)), any given stock’s returns can be explained by some linear combination of a risk free asset return and its beta loading on the market portfolio return. Hence, the factor risk of any stock can be hedged out by shorting its beta loading multiplied by the market factor. However, with the large number of tradable and non-tradable factors that have been documented in the literature (Harvey, Liu, and Zhu (2016)), our question becomes difficult to approach because its answer then significantly depends on which factors the researcher decides to include in his empirical study, and is also affected by the empirical uncertainty with estimating the factor loadings.

The first main contribution of this paper is to *theoretically* argue an effective way to hedge against potentially unidentifiable factor risks of a stock is to answer this dual question: *Given any stock, what portfolio of all other stocks is most similar to it?* Suppose all stocks are exposed to the same set of linear factors but with heterogeneous factor loadings. If one can identify a group of peers that is the most “similar” to a given stock i , then this portfolio is also exposed to similar factor loadings of this stock i . We view this portfolio of peer stocks as the *replicate* of stock i . A long position on stock i and a short position on its replicate will expose the holder to any remaining factor risks of stock i that cannot be completely hedged out by its peer stocks. We show this long-short position exactly equates to the residual factor risks of stock i . Provided an econometrician has a method to find these peer stocks, this long-short position does not require the econometrician to know the true underlying factor structure of the economy. We call the expected returns of this long-short position the *Statistical Arbitrage Risk Premium* (SARP) of stock i , and SARP is the key object of study in this paper.

How do we *empirically* study SARP? Is there a cross-sectional difference in SARP? The second main contribution of the paper is answering these questions. For each month end, we use the *elastic-net* estimator, a machine learning method, to project each stock i ’s past twelve months’ daily returns onto the returns of *every other* stock in the market. The resulting elastic-net projection vector is high-dimensional but very sparse. After a suitable normalization, the projection vector

is then used as investment weights into all stocks other than i . The resulting portfolio is hence a machine learning constructed *replicate* of stock i . As theoretically motivated above, the time-series average return from a long position of stock i and a short position of its replicate is the SARP of stock i . Moreover, we call the elastic-net projection R^2 of each stock i as the *Statistical Arbitrage Risk* (SAR) of stock i ; we say a stock has high SAR if it has a low elastic-net R^2 , while a stock has low SAR if it has high R^2 . The core empirical message of this paper can be succinctly summarized as:

$$\text{SARP is increasing in SAR.}$$

That is in the cross-section, “unique” stocks (i.e. so having low R^2 , and hence high SAR) have a higher SARP than “ubiquitous” stocks (i.e. high R^2 , so low SAR). Over the sample period of January 31, 1976 to December 31, 2020, high SAR stocks have a monthly SARP of 1.368% and low SAR stocks have a monthly SARP of 0.267%, and the difference 1.101% is highly statistically significant. And even without studying SARP, we have the important corollary that high SAR stocks have a monthly return of 1.481% and low SAR stocks have a monthly return of 0.771%, and the difference 0.710% is also highly statistically significant.

Our paper belongs to a growing literature of applying machine learning methods to study empirical asset pricing questions. Broadly speaking, many recent papers in this literature use machine learning methods for factor selection and/or forecasting. We do neither in this paper. There are only two purposes of using a machine learning method in this paper: to identify the SAR of each stock, and to construct the replicate portfolio of each stock. The estimation and inference of SARP for each stock use conventional empirical asset pricing procedures.

Recent papers have applied variants of the *least absolute shrinkage and selection operator* (*LASSO*) estimator of Tibshirani (1996). Feng, Giglio, and Xiu (2020) take advantage of the sparsity property of LASSO and develop a multi-step approach to evaluate the price of risk of a given new factor above and beyond an existing set of factors. Freyberger, Neuhierl, and Weber (2017) uses adaptive group LASSO to select characteristics that provides marginal information for the cross section of expected stock returns. The literature has documented a large set of factors

or characteristics (Harvey, Liu, and Zhu (2016)), and it is hoped that machine learning methods can substantially shrink down the number of factors that can explain the cross-section of returns. Chinco, Clark-Joseph, and Ye (2019) use the LASSO to predict one-minute-ahead return using lagged high frequency returns of other stocks as regressors. Gu, Kelly, and Xiu (2020) apply an extensive battery of machine learning methods and discover that such methods improve predictability accuracy over traditional methods. Shu et al. (2020) is a recent paper that uses an adaptive elastic-net estimator to construct a sparse portfolio that can track a large index portfolio. While both of our papers use the elastic-net estimator, our paper emphasizes the use of this estimator to discover a new asset pricing anomaly, while Shu et al. (2020) emphasizes the use of this estimator to mimic and dimension-reduce a large portfolio.

Our paper is also related to the literature of pairs trading, substitutability of risky assets and statistical arbitrage. Gatev, Goetzmann, and Rouwenhorst (2006) finds pairs of similar stocks using the minimal distance between normalized historical prices, and argue that the resulting pairs trading strategy generates abnormal returns and that the source of this profit is the mispricing of close substitutes. Krauss (2017) is a recent survey of the pairs trading literature. Wurgler and Zhuravskaya (2002) similarly also argue that stocks without close substitutes are likely to have large mispricings; the authors identify a few predefined number of similar stocks using a predefined sorting method. In contrast by using a machine learning method in this paper, the selection of a stock's risky peers is completely data driven. Indeed, the closest substitute of a given stock could potentially be hundreds of all other stocks. Huck (2019) and Avellaneda and Lee (2010) are two recent studies on statistical arbitrage.

Section 1 lays out the theoretical framework of the paper. Section 2 explains our estimation methodology. The main empirical results of the paper are in Section 3. Section 3.6 show additional empirical robustness checks. We conclude in Section 4. We defer the details of the elastic-net procedure to Section A. All proofs to Section 1 are in the Online Supplementary Materials Leung and Tam (2021).

1 Theoretical motivation

We first prove a general theoretical asset pricing result that will guide our empirical research design.

Theorem 1.1. *Suppose there are $N + 1$ risky assets and a single risk-free asset. Assume all of these risky assets are governed by a linear factor structure with K number of factors with risky returns \mathbf{F} ,*

$$R_i = \alpha_i + \boldsymbol{\beta}_i^\top \mathbf{F} + \varepsilon_i \quad (1)$$

and where the idiosyncratic risk ε_i of the i th risky asset is assumed to have zero mean and is independent of \mathbf{F} for all $i = 1, \dots, N + 1$. Suppose there are strictly more risky assets than factors, so $N > K$.

Then the excess returns R_i of any individual risky asset i can be expressed as a linear combination of other risky asset returns as.

$$R_i = \mathbf{b}_i^\top \mathbf{R}_{-i} + \mathbf{a}_i^\top \boldsymbol{\Phi}_i - \mathbf{b}_i^\top \boldsymbol{\varepsilon}_{-i} + \varepsilon_i. \quad (2)$$

That is, the excess returns R_i of any individual asset i can be expressed as a combination of: (i) the $N \times 1$ vector of excess returns of all other of risky assets \mathbf{R}_{-i} ; (ii) the factor loadings on some K risky asset returns $\boldsymbol{\Phi}_i$, and we will call these K assets the factor residuals of asset i ; (iii) the $N \times 1$ vector of idiosyncratic risks $\boldsymbol{\varepsilon}_{-i}$ of all other N risky assets; and (iv) the idiosyncratic risk ε_i of asset i itself.

The $N \times 1$ vector \mathbf{b}_i is dependent on the intercepts $\{\alpha_j\}_{j=1, j \neq i}^{N+1}$ and the entire factor loadings $\{\boldsymbol{\beta}_j\}_{j=1, j \neq i}^{N+1}$ of the economy and whose analytical expression is in the Internet Appendix, and where $\mathbf{a}_i^\top := [\alpha_i, \boldsymbol{\beta}_i^\top]$.

This result tells us given *any* factor structure in the financial markets like (1), of which its theoretical existence can always be justified via Ross (1976), the returns of a single risky asset R_i can be expressed as a linear combination \mathbf{b}_i of returns of *all other* risky assets \mathbf{R}_{-i} like (2). The key intuition of this result is a hedging and replication argument. Suppose the researcher does *not* know the exact identities of the factors $\mathbf{F} = [F_1, \dots, F_K]^\top$ in the economy. But as long as all risky

assets have an exposure to these factors, then any particular risky asset i can use some combination \mathbf{b}_i of other risky stocks to hedge against risky asset i 's factor risks.¹ For the empirical component of our paper, this means projecting one stock's returns onto the returns of all other stocks is not a naive statistical exercise but actually has concrete microeconomic foundations.

The next result will tell us the economic content of Theorem 1.1.

Corollary 1.2. *Suppose the conditions of Theorem 1.1 hold. For each fixed risky asset i , we can find $K + 1$ risky assets whose returns are given by*

$$\Phi_{i,0} = \begin{cases} 1, & \text{if } \alpha_j = 0 \text{ for all } j \neq i \\ 1 - \frac{\boldsymbol{\alpha}_{-i}^\top}{\|\boldsymbol{\alpha}_{-i}\|_2^2} (\mathbf{R}_{-i} - \boldsymbol{\varepsilon}_{-i}), & \text{if otherwise} \end{cases} \quad (3a)$$

$$\Phi_{i,k} = F_k - \mathbf{c}_{i,k}^\top (\mathbf{R}_{-i} - \boldsymbol{\varepsilon}_{-i}), \quad k = 1, \dots, K \quad (3b)$$

where $\mathbf{c}_{i,k}$ is some $N \times 1$ deterministic vector such that only depends on the factor loadings $\{\beta_j\}_{j=1,j \neq i}^{N+1}$ and intercept $\{\alpha_j\}_{j=1,j \neq i}^{N+1}$ structure of the economy (its analytical expression is in the Internet Appendix). Here we denote $\boldsymbol{\alpha}_{-i}^\top := [\alpha_1, \dots, \alpha_{i-1}, \alpha_{i+1}, \dots, \alpha_{N+1}]$ and $\|\cdot\|_2$ is the Euclidean norm on \mathbb{R}^N . We will denote $\boldsymbol{\Phi}_i^\top := [\Phi_{i,0}, \Phi_{i,1}, \dots, \Phi_{i,K}]$.

Corollary 1.2 is what motivates us to refer to those K risky assets with returns $\boldsymbol{\Phi}_i$ as *factor residuals* for asset i . From (2), we see the following decomposition,

$$\mathbf{a}_i^\top \boldsymbol{\Phi}_i = \alpha_i \Phi_{i,0} + \boldsymbol{\beta}_i^\top \boldsymbol{\Phi}_{i,1:K}, \quad (4)$$

where $\boldsymbol{\Phi}_{i,1:K}^\top := [\Phi_{i,1}, \dots, \Phi_{i,K}]$.

The first part $\Phi_{i,0}$ adjusts for any potential mispricings in the economy. As an important special case when there is no pricing at all in the economy² meaning $\alpha_i = 0$ and $\alpha_j = 0$ for all assets $j \neq i$,

¹Some readers may think that the result in Theorem 1.1 and the subsequent Corollary 1.2 can be trivially derived by “inverting the factor loadings” in (1) to derive (2). However, we should note that as there are (significantly) more assets N than the number of factors K in this economy. Thus the economy-wide factor loading matrix is necessarily of dimensions $N \times K$, meaning the matrix is not square and thus not invertible. We take extra care in the proofs to make sure that a proper sense of “factor inversion” is possible in this general setup.

²It is well known that if all risky assets are mean-variance efficient then necessarily there is no mispricing in the economy. See an early discussion of this classical empirical asset pricing test in Black, Jensen, and Scholes (1972).

then $\Phi_{i,0} = 1$ and $\alpha_i \Phi_{i,0} = 0$. If on the other hand, when the intercepts α_i, α_j 's are generically non-zero, then the factor residuals of asset i will explicitly hedge out any level of mispricing in the economy through $\Phi_{i,0}$, and asset i 's net mispricing exposure is $\alpha_i \Phi_{i,0} = \alpha_i - \alpha_i \boldsymbol{\alpha}_{-i}^\top (\mathbf{R}_{-i} - \boldsymbol{\varepsilon}_{-i}) / \|\boldsymbol{\alpha}_{-i}\|_2^2$. Secondly, the next K parts $\Phi_{i,k}$ adjust for factor exposures. From the perspective of an agent who owns asset i , he is exposed to each of the $k = 1, \dots, K$ factor returns through the factor loadings $\boldsymbol{\beta}_i^\top = [\beta_{i,1}, \dots, \beta_{i,K}]$ of (1). However, all the other N risky assets will also have some exposure to the k -th factor. Suppose the agent constructs an “artificial asset” that loads into the return of the k -th factor, while shorting some combination $\mathbf{c}_{i,k}$ of the factor contributions to all the other N risky assets $\mathbf{R}_{-i} - \boldsymbol{\varepsilon}_{-i}$. The resulting artificial asset has the returns $\Phi_{i,k}$. Hence $\Phi_{i,k}$ is precisely the residual exposure of asset i to the k -th factor, after using the returns of all other assets to “hedge” as much as possible this factor risk. The “hedging” nature of these artificial assets motivates us to call them as “factor residuals” for asset i . The holder of asset i is exposed to K number of factor risks, and so he will have to construct K number of these factor residuals. And since asset i has factor exposure to K number of factors, the agent will weigh $\beta_{i,k}$ loadings into the k -th factor residual return $\Phi_{i,k}$, as in the $\boldsymbol{\beta}_i^\top \boldsymbol{\Phi}_{i,1:K}$ term of (4).

Let's rearrange (2) and take its expectation,

$$\mathbb{E}[R_i] - \mathbf{b}_i^\top \mathbb{E}[\mathbf{R}_{-i}] = \mathbf{a}_i^\top \mathbb{E}[\boldsymbol{\Phi}_i] - \mathbf{b}_i^\top \mathbb{E}[\boldsymbol{\varepsilon}_{-i}] + \mathbb{E}[\varepsilon_i] = \mathbf{a}_i^\top \mathbb{E}[\boldsymbol{\Phi}_i], \quad (5)$$

where $\mathbb{E}[\boldsymbol{\varepsilon}_{-i}] = \mathbf{0}_N$ and $\mathbb{E}[\varepsilon_i] = 0$ because idiosyncratic risks are not priced. The overall term $\mathbf{a}_i^\top \mathbb{E}[\boldsymbol{\Phi}_i]$ in (5) is exactly the expected portfolio return into these K factor residuals. This is why we will call $\mathbf{a}_i^\top \mathbb{E}[\boldsymbol{\Phi}_i]$ as the *Statistical Arbitrage Risk Premium (SARP) for asset i*. The key objective of this paper is to empirically study SARP.

The next result shows that, under weak economic and technical conditions, SARP is non-zero.

Corollary 1.3. *The SARP of any non-redundant asset i is almost surely non-zero when some of the K factors are correlated.*

The following result relates the regression R-squared to Theorem 1.1.

Corollary 1.4. Suppose we view (2) as a linear regression of R_i onto the set of regressors \mathbf{R}_{-i} , and \mathbf{b}_i as the vector of regression coefficients. Assume further: (i) $\mathbf{a}_i^\top \mathbb{E}[\Phi_i] \neq 0$; (ii) the idiosyncratic risks are homoskedastic (i.e. $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ if $i \neq j$ and $= \sigma_\varepsilon^2$ if $i = j$); and (iii) the variance-covariances $\text{Var}(\mathbf{F})$ and $\text{Var}(\Phi_i)$ for all i are positive definite. Then,

- (a) The regression R-squared decreases (increases) as $\mathbf{a}_i^\top (\text{Var}(\Phi_i) - \text{Var}(\mathbf{F}))\mathbf{a}_i - \sigma_\varepsilon^2$ becomes more positive (more negative).
- (b) If moreover $\mathbf{a}_i^\top \mathbb{E}[\Phi_i] > 0$, then the regression R-squared is decreasing in $\mathbf{a}_i^\top \mathbb{E}[\Phi_i]$.

In the remainder of this paper, we will identify the regression R^2 with the *Statistical Arbitrage Risk* (SAR) of an asset i . That is, we will say an asset i has a high (low) SAR if it has low (high) R^2 .

Corollary 1.4(a) provides another way to view the K factor residuals of stock i from (3b). Condition (ii) is used to simplify the equations and condition (iii) is a mild technical assumption. For the sake of exposition, consider the case when σ_ε^2 is negligible, and so the magnitude of the term in Corollary 1.4(a) is driven by the positive- or negative-definiteness of the matrix $\text{Var}(\Phi_i) - \text{Var}(\mathbf{F})$. The case where this $K \times K$ matrix is *negative-definite* is when the volatility of the factor residuals of asset i is lower than the volatility of the factors themselves. This happens when the factor residuals of asset i do a good job in hedging asset i against its exposure to the K factor risks. Recall from (3b) these K factor residuals are dependent on the factor structure of all the other N risky assets. This implies these K factor residuals can only do a good job in insuring asset i against factor risks if the N other risky assets also highly co-move with asset i itself, which implies a *high* regression R-squared. Given the desirability of these K factor residuals, the holder of asset i will be willing to pay a high price for these K factor residuals, which then pushes *down* their expected returns. This explains why in Corollary 1.4(b), there is a negative relationship between the regression R-squared and the SARP. The discussion for the case when that $K \times K$ matrix is positive-definite is analogous. The above discussions are still contingent upon the existence and positivity of such a SARP, which again, is entirely an empirical question we now proceed to answer.

2 Empirical hypothesis and methodology

We summarize the empirical implications of our theoretical discussions.

Empirical Hypothesis.

- (1) *The expected difference between a given asset's return and some linear combination of other assets' returns can be seen as a Statistical Arbitrage Risk Premium (SARP). Under weak economic and technical conditions, the SARP is non-zero. Moreover, one does not need to know a priori what are the underlying factors that drive the economy to compute SARP.*
- (2) *We can identify regression R² with Statistical Arbitrage Risk (SAR). We anticipate assets with low SAR (i.e. high R²) to have a low SARP, while assets with a high SAR (i.e. low R²) to have a high SARP in the cross-section.*

The empirical methodology of the paper is separated into two distinct steps. In the first step, we project a given stock's return onto the span of all other stocks' returns to get an empirical approximation of \mathbf{b}_i from (2). However, despite the microfoundations of Theorem 1.1, we shall argue it is econometrically non-trivial to execute this projection. We will apply a machine learning method to overcome a critical technical hurdle. In the second step, we will use standard portfolio sort methods from the empirical asset pricing literature to estimate the expectation $\mathbf{a}_i^\top \mathbb{E}[\Phi_i]$ of (5), which is again the SARP of asset i .

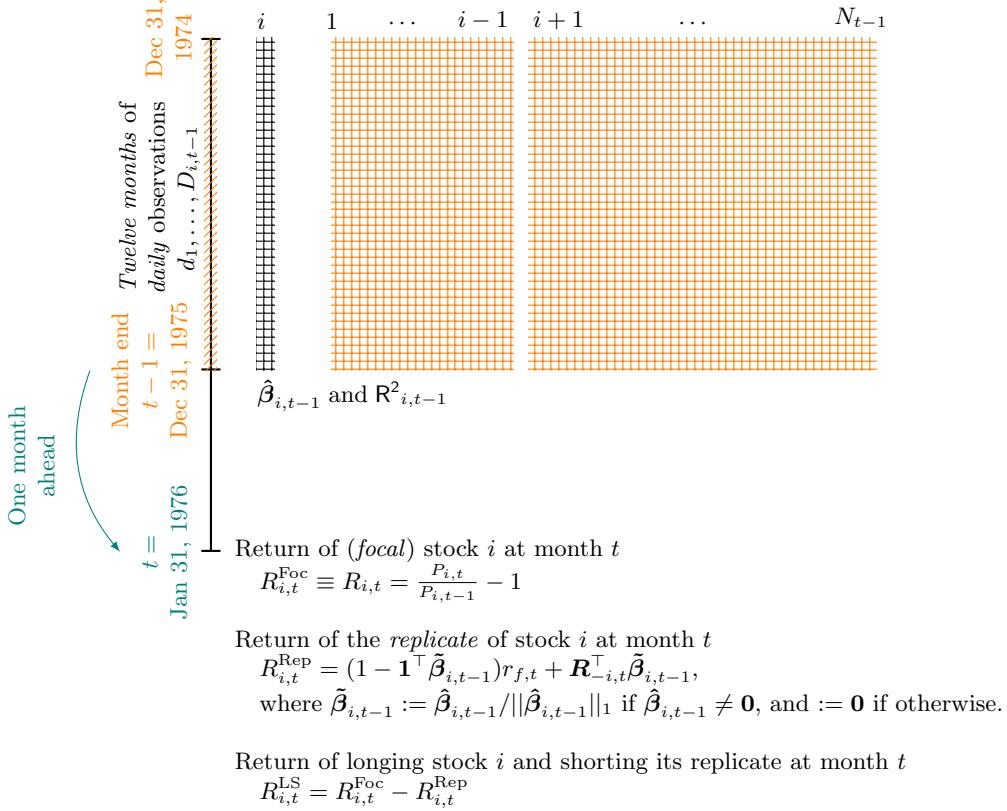
2.1 Data and projection procedure via the elastic-net

2.1.1 Data sources

Our data source is standard. We use both the CRSP daily and monthly data from December 31, 1974 to December 31, 2020 with the standard filters. In all of the subsequent empirical analysis, we identify a stock by its PERMNO number as recorded in the CRSP database.³ Moreover, we only

³As is well known with the CRSP and Compustat databases, there are several unique identifiers of equities: PERMNO, PERMCO and GVKEY. Each have their different strengths and weaknesses. We recognize some — but few — firms have dual class shares which implies a single firm can have having multiple PERMNOs. Rather than making arbitrary corrections to somehow “merge” the time series of returns of these dual class shares, we simply leave the PERMNOs “as is” in CRSP. In all, this means our paper focuses on tracking an individual security rather than the individual firm — although they are synonymous with each other except for those few special cases.

Figure 1: Projection and replicate construction procedure. For each month ending at December 31, 1975, January 31, 1976, ..., November 30, 2020, we use the past twelve months' worth of daily observations to project each of stock returns onto the returns span of every other stock. We only consider stocks have at least 60 days worth for daily trading data. That is, for each stock $i = 1, \dots, N_{t-1}$, suppose $\{d_1, \dots, D_{i,t-1}\}$ with $D_i \geq 60$ are past twelve months' worth of trading days that end at month $t-1$. Note that the number of stocks N_{t-1} in the market at each month end $t-1$ may vary. The returns span for stock i are all those other stocks $j = 1, \dots, i-1, i+1, \dots, N_{t-1}$ that have trading days that at least overlap with that of stock i , so $\{d_1, \dots, D_{i,t-1}\} \subseteq \{d_1, \dots, D_{j,t-1}\}$.



include stocks that, in the past twelve months for any given month end, there are at least 60 days of valid trading returns. This 60 days choice ensures that we do include effectively all stocks, except for the most extremely illiquid or dead stocks, so that our results are not driven only by the liquid and hence most likely large stocks. We also obtain the Fama-French data from Kenneth French's website.

Remark 2.1 (Filters and sampling period). We subset only for US common equities (i.e. SHRCD code of 10 or 11), and only those that are listed in the NYSE, AMEX or NASDAQ (i.e. EXCHCD code of 1, 2 or 3). In addition, we start our data sample from 1974 because the CRSP datasets only started

to include NASDAQ stocks in December 1972. We start in December 1974 to allow an additional year of buffering for good measure. Effectively this means the number of stocks in CRSP pre-1974 and post-1974 are structurally different in several magnitudes; see (Bali et al., 2016, §7.1.2) for a detailed discussion. Our estimation and projection method is clearly sensitive to total number of stocks. Starting our analysis pre-1974 might bias our results simply due to CRSP data limitations.

2.1.2 Projection

The first step in the empirical test of Theorem 1.1 is to project a given stock's returns onto the returns span of all other stocks. Our projection procedure is summarized in Figure 1. For months ending $t - 1 = \text{December 31, 1975, January 31, 1976, ..., November 30, 2020}$, we use the past twelve months' worth of daily observations to project each of stock i 's returns onto the returns of *all other* stocks. We reserve a one month gap for returns realization in a procedure to be described in Section 2.3. Unless specified otherwise, we will denote $t - 1$ as the end of the projection month, and t as the one-month ahead returns realization date; that is, we set $t = \text{January 31, 1976, February 29, 1976, ..., December 31, 2020}$. Figure 3(a) shows the number of stocks that have at least 60 past trading days in for each given month $t - 1$. Observe since we start our projection procedure on December 31, 1975, we require daily data starting from December 31, 1974.

Let N_{t-1} be the total number of stocks traded in the market at month $t - 1$. The return vector of stock i at month $t - 1$ and the returns span of all other assets are, respectively:

$$y_{i,t-1} = \begin{pmatrix} R_{i,d_1} \\ \vdots \\ R_{i,d_{D_{i,t-1}}} \end{pmatrix}_{D_{i,t-1} \times 1} \quad (6a)$$

$$\mathbf{X}_{i,t-1} = \begin{pmatrix} R_{1,d_1} & \dots & R_{i-1,d_1} & R_{i+1,d_1} & \dots & R_{N_{t-1},d_1} \\ \vdots & & \vdots & \vdots & & \vdots \\ R_{1,d_{D_{i,t-1}}} & \dots & R_{i-1,d_{D_{i,t-1}}} & R_{i+1,d_{D_{i,t-1}}} & \dots & R_{N_{t-1},d_{D_{i,t-1}}} \end{pmatrix}_{D_{i,t-1} \times (N_{t-1}-1)} \quad (6b)$$

where $R_{i,d}$ is the daily return of stock i , $D_{i,t-1}$ is the number of trading days of stock i in the

past 12 months ending at month $t - 1$. The dimensions of (6b) are approximately 250×5000 for each stock i . There are $T = 539$ number of months from December 31, 1975 to November 30, 2020. Thus we run a total of approximately $T \times 5000 \approx 2.7$ million projections in this paper.

In this paper we will use the *elastic-net estimator* developed by Zou and Hastie (2005) to empirically project a given stock's return onto the returns span of all other stocks. We defer a detailed and technical discussion of the elastic-net in our context to Section A.1. But if we are interested in estimating *linear* relationships as in Theorem 1.1, why do we not use the workhorse *ordinary least squares* (OLS) estimator? The design matrix (6b) of returns to evaluate our empirical hypothesis is necessarily a $T \times N$ matrix, where $T \approx 250$ is the number of days, and $N \approx 5000$ is the total number of traded stocks. This is a case where $T \ll N$. This means the $N \times N$ matrix $\mathbf{X}_{i,t-1}^\top \mathbf{X}_{i,t-1}$ is *not* full rank. The OLS estimator is thus necessarily *not* well defined. In contrast, the elastic-net is a machine learning method that explicitly allows for “wide” $T \ll N$ regressors.

We denote the elastic-net projection coefficient vector of stock i at month $t - 1$ as $\hat{\beta}_{i,t-1} \in \mathbb{R}^{N_{i,t-1}}$, and the resulting *coefficient of determination* (“*R-squared*”) as $R^2_{i,t-1}$. Note and recall the conventional definition of R^2 is,

$$R^2_{i,t-1} := 1 - \frac{\sum_{s=1}^{D_{i,t-1}} (R_{i,d_s} - \hat{y}_{i,t-1})^2}{\sum_{s=1}^{D_{i,t-1}} (R_{i,d_s} - \hat{\mu}_{i,t-1})^2},$$

where $\hat{y}_{i,t-1} := \mathbf{X}_{i,t-1} \hat{\beta}_{i,t-1}$ is the fitted value, and $\hat{\mu}_{i,t-1}$ is the sample mean; they are explicitly given by

$$\hat{y}_{i,t-1} = \mathbf{X}_{i,t-1} \hat{\beta}_{i,t-1}, \quad \hat{\mu}_{i,t-1} = \frac{1}{D_{i,t-1}} \sum_{s=1}^{D_{i,t-1}} R_{i,d_s}.$$

We emphasize we are only using the elastic-net as a projection method for constructing the replicates, and we do *not* use it for statistical inference. The statistical inference claims are on the expected returns of SARP, which we will discuss beginning in Section 2.3. As a result, despite the large number of projections that we run at this stage, we do not suffer from the multiple hypothesis testing problem that has been discussed in the recent empirical asset pricing literature by Harvey et al. (2016). Other than the use in calculating R^2 , we do *not* use the fitted value $\hat{y}_{i,t-1}$ in any other subsequent steps. This is in contrast to the recent literature on financial applications of machine

learning (say Gu et al. (2020) and others) where they use the fitted value (or predicted value when one uses $\mathbf{X}_{i,t}$ instead of $\mathbf{X}_{i,t-1}$) in assessing model forecasting accuracy.

Remark 2.2 (Not projecting onto the intercept). We deliberately do *not* include an intercept as a regressor in (6b). Including an intercept together with the sparseness property of the elastic-net estimator will attribute a stock with stale prices with extremely high R^2 . See Section A.1.2 for a more detailed discussion. Other than the projection method here and in Section 3.6.2, unless noted otherwise, all subsequent statistical inference tests that employ a linear regression will include an intercept.

Remark 2.3 (Why elastic-net and not other machine learning methods?). Out of a myriad of machine learning methods, why did we choose the elastic-net estimator to test our empirical implication? Simply put, we regard the elastic-net estimator as a parsimonious method that allows us to test our theoretical prediction. We defer to Section A.1 for detailed technical discussions of why we particularly use and prefer the elastic-net estimator in this paper.

Remark 2.4 (Overlapping data). It is evident we are using overlapping time data. Overlapping returns data in traditional empirical asset pricing raises several technical inference issues revolving around autocorrelations (see Hansen and Hodrick (1980) and a recent discussion by Hedegaard and Hodrick (2016)). However, the “wide” regressors in our setting imply we actually gain ≈ 5000 new cross-sectional data points for each month advance. Because of the sheer amount of new entering cross sectional data, it is not necessarily true that the projected coefficient ending at months $t - 2$ and $t - 1$ would be quantitatively similar even if they only differ by a one month step.

2.2 Replicate construction

For each month end $t - 1$ and each stock $i = 1, \dots, N_{t-1}$, we collect the projected coefficient $\hat{\beta}_{i,t-1}$ and the $R^2_{i,t-1}$. We track stock i ’s *one-month ahead* return $R_{i,t}$ from $t - 1$ to t . For subsequent exposition clarity, we will sometimes call stock i as the *focal stock* and write its return at month t as,

$$R_{i,t}^{\text{Foc}} \equiv R_{i,t}. \quad (7)$$

Next, we introduce the *replicate* (portfolio) of a stock i . This is a key idea of this paper. Let $\mathbf{R}_{-i,t} := [R_{1,t}, \dots, R_{i-1,t}, R_{i+1,t}, \dots, R_{N_{t-1},t}]^\top$ be the vector of month t returns for all stocks except stock i . We wish to treat the projected coefficients $\hat{\beta}_{i,t-1}$ as investment weights into each of the $N_{t-1} - 1$ number of stocks. However, the regularization nature of the elastic-net causes the entries of these estimated coefficients to be small in magnitude. So if we were to directly use $\hat{\beta}_{i,t-1}$ as investment weights, the result would be a very small allocation into risky component of the portfolio. To have a more reasonable risky component, we normalize $\hat{\beta}_{i,t-1}$ by its L^1 norm and write,⁴

$$\tilde{\beta}_{i,t-1} := \begin{cases} \mathbf{0} & \text{if } \hat{\beta}_{i,t-1} = \mathbf{0}, \\ \frac{\hat{\beta}_{i,t-1}}{\|\hat{\beta}_{i,t-1}\|_1} & \text{otherwise.} \end{cases} \quad (8)$$

Since investment weights must sum to one, we place the remainder of the weights $\mathbf{1}^\top \tilde{\beta}_{i,t-1}$ into the risk free asset with returns $r_{f,t}$, and here $\mathbf{1}$ is a vector of ones of conformable dimensions. In all, the month t return of stock i 's *replicate* is,

$$R_{i,t}^{\text{Rep}} := (1 - \mathbf{1}^\top \tilde{\beta}_{i,t-1})r_{f,t} + \mathbf{R}_{-i,t}^\top \tilde{\beta}_{i,t-1}. \quad (9)$$

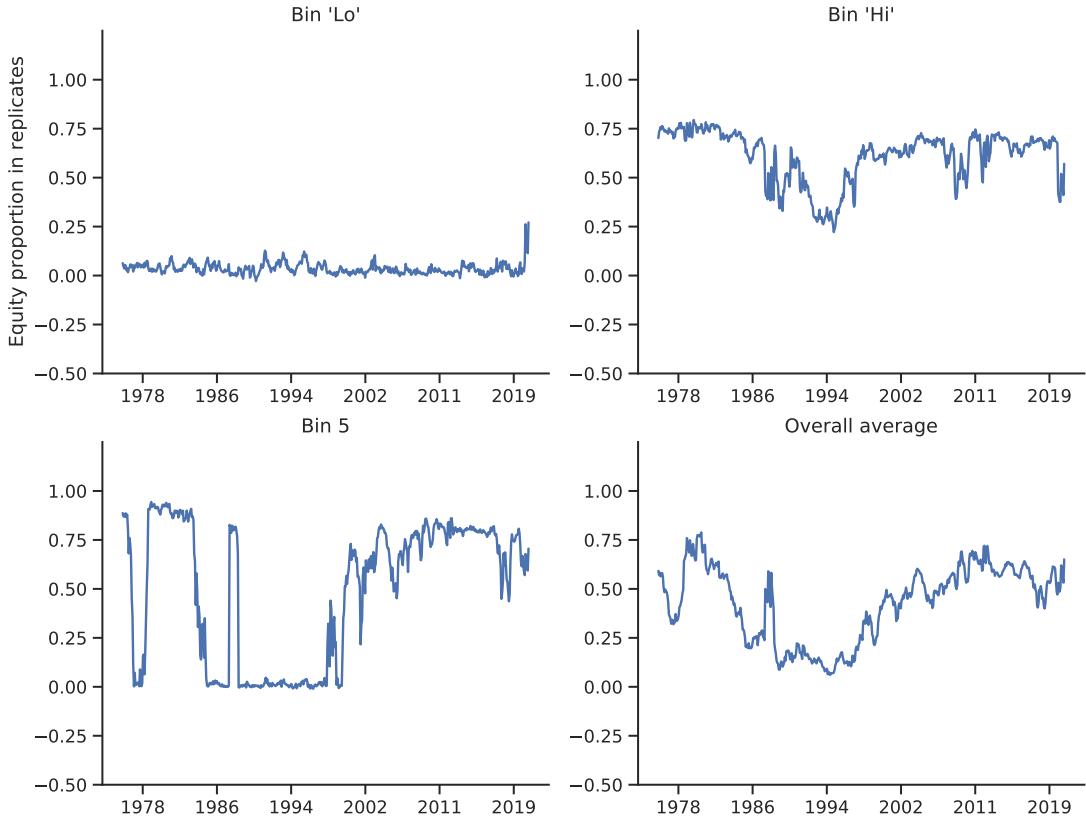
Finally, we track the *long-short* return of stock i against its replicate,

$$R_{i,t}^{\text{LS}} := R_{i,t}^{\text{Foc}} - R_{i,t}^{\text{Rep}}. \quad (10)$$

This long-short return (10) will proxy for the difference $R_i - \mathbf{b}_i^\top \mathbf{R}_{-i}$ in Theorem 1.1, and is the key emphasis of study in this paper. We will use conventional portfolio sort methods of the empirical asset pricing literature to estimate the expectation $\mathbb{E}[R_i] - \mathbf{b}_i^\top \mathbb{E}[\mathbf{R}_{-i}]$, which then is equal to the

⁴For any vector $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, we define its L^1 norm as $\|\mathbf{x}\|_1 := \sum_{j=1}^n |x_j|$. Note that the L^1 norm shows up explicitly in the objective function of the elastic-net estimator. The L^2 norm also shows up in the elastic-net objective function; as is usual, $\|\mathbf{x}\|_2 := \sqrt{\sum_{j=1}^n x_j^2}$. However, we choose to normalize by L^1 and not L^2 because of scaling. We have inequality $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$. Economically and empirically, this implies scaling by the L^2 norm will result in a highly levered equity component for the replicate for stock i . The L^1 scaling generates more a more moderate result.

Figure 2: Proportion of equity invested in the replicates. We consider the average amount of equity-only components that are invested by each replicate portfolio. Stocks are sorted into their method m R^2 decile bins $k = \text{'Lo'}, \dots, \text{'Hi'}$. At the end of the estimation month $t - 1$, we have the normalized vector $\tilde{\beta}_{i,t-1}$ of asset i . The quantity $\mathbf{1}^\top \tilde{\beta}_{i,t-1}$ is the proportion allocated to the equity-only components of stock i 's replicate as from (9). The average equity-only proportion across all the replicates in bin k at month $t - 1$ is the quantity $\frac{1}{|B_{t-1}^k|} \sum_{i \in B_{t-1}^k} \mathbf{1}^\top \tilde{\beta}_{i,t-1}$. Here we plot time series of this average equity-only proportion for the bins $k = \text{'Lo'}, 5$ and 'Hi' , and the y -axis are expressed in decimals (e.g. 0.10 means 10%). The plot labeled “Overall average” plots the overall average of these equity-only components across all bins, which is the quantity $\frac{1}{10} \frac{1}{|B_{t-1}^k|} \sum_{k=1}^{10} \sum_{i \in B_{t-1}^k} \mathbf{1}^\top \tilde{\beta}_{i,t-1}$. The sampling period is from December 1975 to November 2020.



theoretically motivated SARP $\mathbf{a}_i^\top \mathbb{E}[\Phi_i]$ of stock i .

Remark 2.5 (Projection method for constructing mimicking portfolios). Breeden et al. (1989) and Lamont (2001) use analogous methods to (9) to construct a mimicking factor out of tradable base assets. Ang et al. (2006) use also an analogous method to construct a factor mimicking aggregate volatility risk. While our approach in (9) seems identical to these previous methods, we stress the dimensionality of the regressors is substantially different. The number of base assets in those aforementioned methods is small, usually in the range of five to ten. In contrast, our base assets are effectively every other stock other than the stock i itself, which number in the thousands. Wurgler and Zhuravskaya (2002)⁵ uses an idea that is conceptually similar — but operationally quite different — to this paper in evaluating arbitrage risk. Wurgler and Zhuravskaya (2002) first use a sorting method to closely match a stock i with three other stocks that is most closest to it on size and book-to-market, and then use a linear regression on of the returns of stock i onto these three stocks. Similar to us, they use the regression coefficients to construct a replicate of stock i . However, the selection procedure of Wurgler and Zhuravskaya (2002) by portfolio sort is not entirely statistical in nature. Indeed, a priori there's no reason why the time series statistical properties of any given stock i would be best matched by the top three stock matched on size and value, even if size and value are well known priced factors. Moreover, the focus of Wurgler and Zhuravskaya (2002) is to study “demand curves” for stocks and to identify this effect, the authors only constrain their study to 259 stocks addition into the S&P 500 index over a 13 year period. In contrast, the focus of our study is SARP and we consider effectively all stocks (so not just those in the S&P 500) at any given point in time and over a 45 year period.

2.3 Portfolio formation and sort

Having defined three types of returns (7), (9) and (10) associated with stock i , we now proceed to construct portfolios. We use our theoretical discussions as guidance: Corollary 1.4 anticipates we should find a negative relationship between the regression R-squared (the SAR of stock i) and the SARP of stock i . At the end of the month $t - 1$, we sort each stock i by its $R^2_{i,t-1}$ into *decile* bins.

⁵We are grateful for an anonymous referee for pointing out this useful reference to us.

Let B_{t-1}^k be the set of stocks in the k -th bin at month $t - 1$. We organize the bins in ascending order, so bin $k = 1$ (labeled “Lo”) consists of stocks with the lowest R^2 ’s, and bin $k = 10$ (labeled “Hi”) consists of stocks with the highest R^2 ’s.

We focus on equal-weighted portfolios in this paper (see Remark 2.6 for a discussion of the issues in evaluating value-weighted SARP). The equal-weighted *excess returns* of bin $k = \text{Lo}, 2, \dots, \text{Hi}$ of the stocks is standard:

$$\bar{R}_{\text{Foc},t}^k = \frac{1}{|B_{t-1}^k|} \sum_{i \in B_{t-1}^k} (R_{i,t} - r_{f,t}), \quad (11)$$

where we denote $|B|$ as the cardinality of a set B . In this paper, we will view and assume the replicate of stock i to have the same equal-weight as its focal counterpart,

$$\bar{R}_{\text{Rep},t}^k := \frac{1}{|B_{t-1}^k|} \sum_{i \in B_{t-1}^k} (R_{i,t}^{\text{Rep}} - r_{f,t}). \quad (12)$$

Finally, the construction of the equal-weighted portfolio of the long-short of the focal stocks versus their replicates is now immediate thanks to the aforementioned equivalent weighting assumption. We define the equal-weighted portfolio return of the *long-focal, short-replicate position* in bin k as,

$$\bar{R}_{\text{LS},t}^k := \bar{R}_{\text{Foc},t}^k - \bar{R}_{\text{Rep},t}^k = \frac{1}{|B_{t-1}^k|} \sum_{i \in B_{t-1}^k} R_{i,t}^{\text{LS}}. \quad (13)$$

Note (13) does not subtract a risk-free rate as it is the difference of two excess returns.

Remark 2.6 (Ambiguity in value-weighted SARP). Value-weighting causes several causes for concern in evaluating SARP. Value-weights for focal stocks in each bin k is conventional. The cause for caution is what should one assume for the value-weights of the replicates. Perhaps the most natural way is just assume the replicate of stock i has the same value-weight as stock i itself. While this value-weight assumption for the replicates might seem natural, it is also ambiguous. For stocks with high SAR there is no contention: for these high SAR stocks, their replicates are effectively just the risk-free asset, and so using value-weights in this way would just produce the equity risk premium. But for stocks with low SAR, there will be many similar peer stocks in their replicates.

It is possible that a focal stock has large size, but the elastic-net statistically matches its peers with small stocks, and vice-versa. Thus the total value of the replicate portfolio of stock i could be substantially different than that of the focal stock i itself. Assigning the replicate of stock i to have the same value-weight of the focal stock i in bin k may or may not reflect the size of the latter. Indeed, the same concern also holds in converse. This ambiguity in what “value-weighting” means for the replicates is what drives our paper to focus on the most parsimonious weighting scheme — equal-weights.

3 Empirical results

We can now translate our theoretical predictions of Section 1 to concrete quantities for an empirical investigation: (i) By Corollary 1.3, SARP should be non-zero (and indeed positive). That is, we should expect the time-series average of $\bar{R}_{LS,t}^k$ of (10) to be non-zero in the data; and (ii) More importantly by Corollary 1.4, SARP should be increasing with SAR. In other words, stocks that have high SAR (low R^2) should have high SARP, and stocks with low SAR (high R^2) should have low SARP.

3.1 Summary statistics

3.1.1 Characteristics summary statistics

Let’s first investigate the average characteristics of stocks that are decile sorted by elastic-net R^2 ’s. Let $Z_{i,t-1}$ be some scalar characteristic of stock i at month $t - 1$. For each bin k , we compute the simple average of stocks’ characteristics $\bar{Z}_{t-1}^k := \frac{1}{|B_{t-1}^k|} \sum_{i \in B_{t-1}^k} Z_{i,t-1}$. The numbers in Table 1 show the time-series mean, standard deviation, 5th and 95th percentiles of these characteristics \bar{Z}_{t-1}^k for each bin k .

Result (i) of Table 1 shows the elastic-net R^2 characteristic $Z_{i,t-1} = R^2_{i,t-1}$. The elastic-net is capturing a wide range of projection R^2 ’s from effectively 0% in the lowest bin to 62% in the highest bin.⁶ As a matter of comparison and for subsequent robustness discussions in Section 3.6.2, at the

⁶While there is some resemblance, we caution that the positive association of R^2 and number of non-zero elements in the estimated elastic-net coefficient vector is different from the OLS case where including more regressors necessarily

end of month $t - 1$ we take each stock i 's past twelve months' of daily returns and project them onto the daily returns of the Fama and French (2015) five factors using OLS. We collect the resulting FF5 OLS R^2 and this is also a characteristic for each stock i for month $t - 1$. Result (ii) shows the summary statistics of the FF5 OLS R^2 characteristic for stocks sorted by elastic-net R^2 . It is quite interesting to observe that stocks with high SAR (low elastic-net R^2) have also low corresponding FF5 OLS R^2 , and likewise stocks with low SAR (high elastic-net R^2) have high FF5 OLS R^2 .

Results (iii) and (iv) examine the quality of our replicate construction procedure. Recall (9). The characteristic $Z_{i,t-1} = \text{Number of non-zero entries in } \hat{\beta}_{i,t-1}$ is the number of risky peers used to construct the replicate of stock i and (iii) shows this result. Note the sparseness of the elastic-net projection vector: while N_{t-1} ranges in the thousands, the number of non-zero entries in the projection vector is remarkably low. For stocks with the highest SAR, it is essentially so “unique” that the replicate of this stock is just the risk-free asset. In other words, the next best alternative of not investing into a truly unique stock is simply the risk-free asset. In contrast for stocks with the lowest SAR, one can statistically identify about 82.58 risky peers. So even if one forgoes investing into a “ubiquitous” stock, one can construct its replicate consisting of 82.58 peer stocks that have similar statistical properties. Next, the characteristic $Z_{i,t-1} = \mathbf{1}^\top \tilde{\beta}_{i,t-1}$ is the proportion of equity that is used to construct the replicate of stock i . Result (iv) shows we allocate only 4% into equity components (so 96% into the risk-free asset) of the replicates for stocks with the highest SAR, but we allocate on average 60% into equity components of the replicates for stocks with the lowest SAR. At least with our elastic-net projection procedure and even for stocks with the highest SAR, we basically never achieve an 100% allocation into equity for the replicates of any stock; an 100% allocation can happen if there is a stock l physically different from stock i but are otherwise statistically equivalent. That is to say, *all* stocks are effectively unique but there's still nonetheless a wide range of heterogeneity.

Results (v) to (vii) show the market capitalization, book-to-market and dollar volume liquidity characteristics. From result (v), low SAR stocks tend to be smaller stocks while high SAR stocks tend to be bigger stocks. Nonetheless, we should keep in mind the magnitudes and also within and mechanically raises R^2 . In our elastic-net application, the number of regressors N_{t-1} remains *fixed*, and it is an estimation outcome that only few of them have non-zero coefficient loadings.

decile bin heterogeneity. For low SAR stocks the average market capitalization is about \$1.2 billion with a standard deviation of \$9.3 billion, and for the high SAR stocks the average is \$6.1 billion with a standard deviation of \$26.2 billion. In all, while the high SAR bin holds on average “mega-sized” stocks, the stocks in the low SAR bin are not exclusively microcap stocks either. Hence our subsequent cross-sectional results are unlikely to be driven by a size effect. Result (vi) shows an interesting value tilt for stocks with high SAR and a growth tilt for stocks with low SAR. Finally, result (vi) shows the dollar volume liquidity (VOLD) characteristic. In particular, $VOLD_{i,t-1}$ is defined as the product of trading volume of stock i on the last day of month $t - 1$ and the closing price of stock i on the last day of month $t - 1$, then all divided by 1,000,000. We see that on average, high SAR stocks have a lower liquidity than low SAR stocks. On the surface, results (v) to (vii), it appears size, book-to-market and liquidity are related to a stock’s SAR. We will explicitly control for these three characteristics, among others, in Section 3.4 to ensure our results are not driven by the well-known pricing qualities of these characteristics in the cross-section.

Pursuant to Table 1(i), we directly plot the time series of the cross-sectionally averaged R^2 characteristic in Figure 3(b). Again, as motivated by our theoretical discussions, we regard assets with low R^2 to have high SAR, and assets with high R^2 to have low SAR. As a consequence, the plots in Figure 3(b) can thus be regarded as an “average” SAR of the economy at any given point in time. By inspection, it appears the average SAR of the economy is countercyclical and spikes considerably during times of financial distress (e.g. 1980-1991 US savings and loan crisis, 1998 Russian financial, 2008 - 2009 Great Recession, Greek government debt crisis 2011-2012, March 2020 COVID-19 crash, and others). These plots strongly suggest SAR (and the resulting SARP of assets) is not simply a statistical construct but is also capturing systematic shocks. We formally test the association of SAR with macroeconomic variables in Figure 3(c). The regression results corroborate the visual inspection of Figure 3(b): shocks to the average SAR is countercyclical, in that it is negatively associated with shocks to both personal consumption expenditure and consumer sentiment.

Table 1: Summary characteristics of elastic-net R² sorted stocks. We project each stock's past twelve month's daily returns onto every other stock using the elastic-net estimator. Stocks are sorted into deciles based on their elastic-net R² from the lowest (quantile 1, labelled "Lo") to highest (quantile 10, labelled "Hi"). At the end of each month, we first compute the simple average of the stocks' characteristics within each bin. We then take the time-average of these averaged characteristics for each bin, and the displayed figure shows this cross-sectional and time-series average for each bin. Brackets show the standard deviation and the parentheses show the (5th, 95th) percentiles of the characteristics. The sampling period is from December 31, 1975 to November 30, 2020.

	EN R ² Lo	2	3	4	5	6	7	8	9	EN R ² Hi
(i) EN R ²	-0.01 [0.01]	-0.00 [0.02]	0.02 [0.07]	0.07 [0.13]	0.12 [0.18]	0.18 [0.22]	0.25 [0.25]	0.33 [0.28]	0.42 [0.29]	0.62 [0.28]
(ii) FF5 OLS R ²	(-0.03, -0.00) 0.08 [0.07]	(-0.01, 0.01) 0.07 [0.07]	(-0.00, 0.14) 0.09 [0.09]	(-0.00, 0.37) 0.11 [0.11]	(-0.00, 0.52) 0.14 [0.13]	(-0.00, 0.62) 0.17 [0.15]	(0.00, 0.71) 0.21 [0.17]	(0.01, 0.79) 0.25 [0.19]	(0.03, 0.86) 0.31 [0.21]	(0.11, 0.98) 0.31 [0.21]
(iii) Non-zero entries of EN proj vector	(0.01, 0.22) 0.50 [0.60]	(0.01, 0.20) 0.72 [1.63]	(0.01, 0.21) 2.34 [5.73]	(0.01, 0.27) 6.17 [11.66]	(0.01, 0.35) 11.37 [17.00]	(0.01, 0.42) 17.76 [21.89]	(0.01, 0.48) 25.22 [26.84]	(0.01, 0.54) 34.07 [32.28]	(0.02, 0.60) 46.85 [38.45]	(0.02, 0.68) 82.58 [54.92]
(iv) Sum of EN proj vector entries	(0.00, 1.00) 0.04 [0.26]	(0.00, 2.00) 0.07 [0.33]	(0.00, 12.00) 0.20 [0.45]	(0.00, 32.00) 0.33 [0.51]	(0.00, 46.00) 0.44 [0.51]	(0.00, 59.00) 0.50 [0.51]	(1.00, 74.00) 0.58 [0.58]	(1.00, 92.00) 0.61 [0.50]	(3.00, 115.00) 0.62 [0.38]	(10.00, 186.00) 0.60 [0.26]
(v) MktCap (in mil)	(0.00, 0.93) 1,214.89 [9,253.69]	(0.00, 1.00) 863.83 [7,449.02]	(0.00, 1.00) 888.31 [8,222.65]	(-0.08, 1.00) 1,262.53 [11,497.41]	(-0.04, 1.00) 1,548.21 [12,337.95]	(-0.20, 1.00) 1,817.25 [13,904.73]	(-1.00, 1.00) 2,223.06 [16,214.65]	(-0.81, 1.00) 2,681.28 [16,726.44]	(-0.22, 1.00) 3,450.98 [19,528.92]	(0.11, 0.91) 6,109.52 [26,175.47]
(vi) Book-to-market	(5.55, 3,644.49) 0.95 [1.17]	(4.33, 2,283.37) 0.98 [1.16]	(3.99, 2,097.74) 0.96 [1.17]	(4.13, 2,828.01) 0.90 [1.14]	(4.59, 4,027.00) 0.85 [1.07]	(5.27, 5,272.39) 0.81 [0.98]	(6.47, 6,889.35) 0.78 [0.94]	(7.71, 9,005.63) 0.75 [0.87]	(9.48, 12,596.65) 0.72 [0.84]	(14.86, 23,868.21) 0.68 [0.68]
(vii) Dollar volume liquidity (in mil)	(0.11, 2.45) 6.96 [56.09]	(0.11, 2.57) 4.67 [37.73]	(0.10, 2.54) 5.46 [50.05]	(0.10, 2.38) 8.71 [107.06]	(0.09, 2.24) 11.39 [107.76]	(0.09, 2.12) 13.70 [109.11]	(0.10, 2.01) 16.80 [114.71]	(0.10, 1.89) 20.39 [117.61]	(0.10, 1.79) 26.66 [127.33]	(0.10, 1.64) 44.11 [166.52]
	(0.01, 23.11) [0.01, 23.11]	(0.01, 13.79) [0.01, 13.79]	(0.01, 14.42) [0.01, 14.42]	(0.01, 24.60) [0.01, 37.15]	(0.01, 37.15) [0.01, 50.20]	(0.01, 50.20) [0.01, 64.98]	(0.01, 64.98) [0.02, 86.54]	(0.02, 123.23) [0.02, 123.23]	(0.03, 198.25) [0.03, 198.25]	

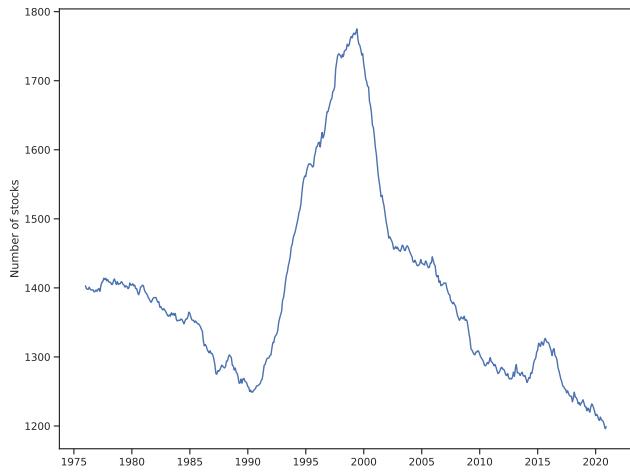
3.1.2 Time series correlations and replicates as hedging instruments

How good are the replicates in mimicking the return behavior of the focals? While we will discuss at length the first moments of the focals and the replicates starting from Section 3.2, let's first investigate into their second cross-moments. For bins $k_1, k_2 \in \{\text{Lo}, 2, \dots, \text{Hi}\}$ and return type $m_1, m_2 \in \{\text{Foc}, \text{Rep}\}$ (recall (11) and (12)), we track the time series $\bar{R}_{m_1,t}^{k_1}$ and the time series $\bar{R}_{m_2,t}^{k_2}$. The numbers in Table 2 reports the sample correlation between the time series $\bar{R}_{m_1,t}^{k_1}$ and $\bar{R}_{m_2,t}^{k_2}$. The diagonal entries are most important: stocks with low (high) R^2 's have low (high) correlations with their replicates. Specifically, the correlation between the returns of stocks with highest SAR (so they have the lowest R^2) with their replicates is only 17.8%. In contrast, this correlation rises substantially along the diagonal of Table 2 to 91.9% for stocks with the lowest SAR (so highest R^2). In all, for stocks with low SAR, the elastic-net replicate portfolio procedure (that is entirely based on past returns) do a fairly good job in matching the stock's one-month ahead return second moments. However, but expectedly, the replicates do a poor job in mimicking stocks with high SAR.

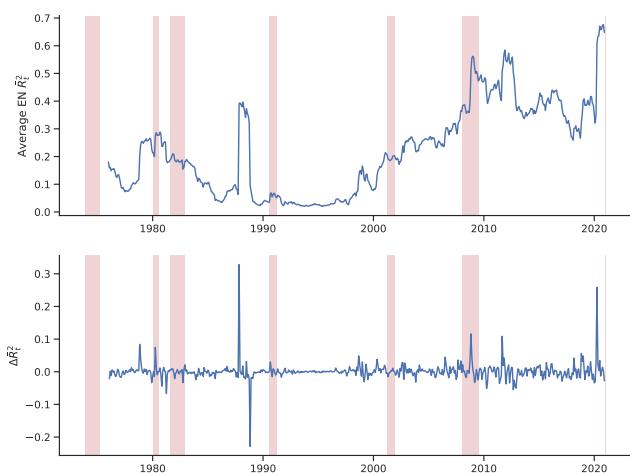
Table 2: Pairwise correlations of one-month ahead returns between stocks and their replicates. The (k_1, k_2) -th entry above represents the correlation of the time series one month ahead returns between the k_1 -th elastic-net R^2 sorted bin of the stocks and the k_2 -th portfolio bin its replicate. The entries are expressed in decimals (e.g. 0.10 means 10%). We form equal-weighted portfolios decile portfolios every month by projecting each stock's daily return over the past year onto every other stock using the elastic-net estimator. Stocks are sorted into deciles based on their elastic-net R^2 from the lowest (quantile 1, labelled “Lo”) to highest (quantile 10, labelled “Hi”). The rows labelled “Foc” report the portfolio of the focal stocks (7). The columns labelled “Rep” report the replicate returns of the focal stock that are constructed out of the estimated elastic-net beta coefficients according to (9). The sampling period is from January 31, 1976 to December 31, 2020.

EN R^2	Lo Rep	2	3	4	5	6	7	8	9	Hi Rep
Lo Foc	0.178	0.358	0.500	0.613	0.661	0.705	0.693	0.678	0.729	0.801
2	0.208	0.363	0.499	0.642	0.709	0.744	0.738	0.724	0.770	0.826
3	0.210	0.416	0.542	0.664	0.735	0.762	0.757	0.742	0.782	0.839
4	0.220	0.395	0.543	0.695	0.762	0.780	0.778	0.764	0.803	0.857
5	0.187	0.370	0.525	0.688	0.769	0.790	0.782	0.769	0.811	0.868
6	0.188	0.337	0.491	0.664	0.758	0.793	0.798	0.789	0.836	0.893
7	0.189	0.328	0.473	0.649	0.763	0.813	0.813	0.807	0.855	0.904
8	0.185	0.306	0.450	0.629	0.760	0.819	0.828	0.830	0.879	0.918
9	0.182	0.289	0.426	0.601	0.737	0.806	0.818	0.824	0.884	0.927
Hi Foc	0.166	0.275	0.416	0.573	0.684	0.752	0.759	0.765	0.833	0.919

Figure 3: Number of regressors, average elastic-net R^2 , and macro regressions. Figure (a) shows the time series of the total number of traded stocks N_{t-1} in the US at any given point in time. The elastic-net R^2 for each stock i at month $t - 1$ is denoted as $R^2_{i,t-1}$. The top panel of Figure (b) plots the time series of the averaged R^2 across all stocks \bar{R}^2_t , and the bottom panel plots the difference $\Delta \bar{R}^2_t := \bar{R}^2_t - \bar{R}^2_{t-1}$. Table (c) regresses the time series $\Delta \bar{R}^2_t$ onto the following regressors' difference and its lag: industrial production, personal consumption expenditure, unemployment rate, University of Michigan consumer sentiment, and the market factor. All macro variables data are from FRED Economic Data of the St. Louis Federal Reserve. Since UMCSENT is only readily available from January 1978, the regressions' sampling period is monthly from January 31, 1978 to December 31, 2020.



(a) Dimensionality value N_{t-1}



(b) Average EN R^2 and lagged its difference

	(1)	(2)	(3)
	EN $\Delta \bar{R}^2_t$		
$\Delta \log \text{INDPRO}_t$	0.092 (0.278)	0.305 (0.951)	0.229 (0.871)
$\Delta \log \text{PCE}_t$	-0.820 (-2.138)	-1.082 (-2.074)	-0.893 (-2.044)
ΔUNRATE_t	-0.750 (-0.879)	-0.980 (-1.118)	-0.505 (-0.694)
$\Delta \log \text{UMCSENT}_t$	-0.081 (-3.079)	-0.075 (-3.294)	-0.047 (-2.262)
$\Delta \log \text{INDPRO}_{t-1}$		-0.289 (-1.195)	-0.111 (-0.587)
$\Delta \log \text{PCE}_{t-1}$		-0.046 (-0.148)	0.075 (0.270)
$\Delta \text{UNRATE}_{t-1}$		0.110 (0.224)	0.472 (1.008)
$\Delta \log \text{UMCSENT}_{t-1}$		-0.031 (-1.617)	-0.029 (-1.592)
MktRF_t			-0.183 (-2.746)
Constant	0.002 (1.431)	0.003 (1.519)	0.004 (1.918)
Observations	514	513	513
R^2	0.073	0.089	0.176
Adjusted R^2	0.065	0.075	0.161

(c) Regressing the difference of EN R^2 on macro variables

3.2 Main empirical result: Statistical arbitrage risk premium (SARP)

The univariate portfolio sorts of Table 3 show the main empirical result of this paper. Let's first discuss the returns of focal stocks in the first column of Table 3. Stocks with the *lowest* elastic-net R^2 earn an excess return of 1.481% (t -stat 5.298) per month, while stocks with the *highest* elastic-net R^2 earn an excess return of 0.771% (t -stat 2.882). The *difference* in returns between stocks with the lowest and highest elastic-net R^2 is 0.710% (t -stat 4.386). In other words, high SAR stocks have a substantially greater return than low SAR stocks.

Table 3: (MAIN RESULTS) Univariate portfolio sort by elastic-net R^2 . We form equal-weighted decile portfolios every month by projecting each stock's past twelve month's daily returns onto every other stock using the elastic-net estimator. Stocks are sorted into deciles based on their elastic-net R^2 from the lowest (quantile 1, labelled "Lo") to highest (quantile 10, labelled "Hi"). The row labelled "Lo - Hi" is the monthly return difference between the "Lo" bin and the "Hi" bin. The row labelled "Avg" is the simple average of the monthly excess returns across the ten $k = \text{'Lo}', 2, \dots, \text{'Hi'}$ bins. The column labelled "Focal" reports the one-month ahead portfolio excess returns (7). The column labelled "Replicate" reports the one-month ahead excess returns of a portfolio that are constructed out of the estimated normalized elastic-net coefficients according to (9). The column labelled "Foc - Rep" reports the one-month ahead returns of the portfolio of a long position in the focal stocks, and a short position in the corresponding replicates. We define the *Statistical Arbitrage Risk Premium* (SARP) as the returns from "Foc - Rep". In addition, a stock with low EN R^2 is said to have high *Statistical Arbitrage Risk* (SAR), while a stock with high EN R^2 is said to have low SAR. The "mean" column is reported in monthly percentage terms (e.g. 1.0 means 1%). Robust Newey and West (1987) t -statistics with six lags are reported in column " t " in parentheses. The sample period is monthly from January 31, 1976 to December 31, 2020.

EN R^2	Focal		Replicate		Foc - Rep	
	mean	t	mean	t	mean	t
Lo	1.481	(5.298)	0.115	(2.109)	1.368	(4.929)
2	1.167	(4.047)	0.346	(2.767)	0.818	(3.056)
3	1.008	(3.408)	0.383	(1.516)	0.633	(2.459)
4	0.941	(3.186)	0.432	(1.451)	0.511	(2.280)
5	0.915	(3.196)	0.360	(1.102)	0.561	(2.634)
6	0.889	(3.196)	0.465	(1.400)	0.431	(2.089)
7	0.869	(3.177)	0.724	(2.080)	0.149	(0.720)
8	0.891	(3.198)	0.792	(2.303)	0.093	(0.470)
9	0.813	(2.943)	0.655	(2.130)	0.161	(1.112)
Hi	0.771	(2.882)	0.503	(2.070)	0.267	(2.581)
Lo - Hi	0.710	(4.386)	-0.387	(-1.662)	1.101	(4.132)
Avg	0.974	(3.584)	0.478	(2.069)	0.499	(3.391)

By conventional wisdom in the empirical asset pricing literature, these results already strongly hint that R^2 is a potential priced factor in the cross-section. However, guided by our theoretical discussions of Section 1, we are not just interested in testing the cross-sectional difference in the

focal stocks (but see later in Section 3.5 for a factor discussion). Rather, we are interested in empirically testing for the cross-sectional presence of SARP. Let's consider the empirical results of the replicates in the second column of Table 3. The excess returns of the replicates are almost monotonically increasing, from 0.115% (*t*-stat 2.109) per month in the *lowest R*² bin, to 0.503% (*t*-stat 2.070) in the *highest R*² bin, even though their *difference* -0.387% (*t*-stat -1.662) is only weakly statistically significant. The excess returns of the replicates in some middle bins (e.g. bins 3 - 6) are statistically insignificant from zero. Other than these four middle bins, it appears the projection construction procedure does construct a replicate asset that has reasonable mean returns.

Finally, we come to the main highlight of our paper: the third column of Table 3 proxies for the SARP. The focals minus replicates have an average return of 1.368% (*t*-stat 4.929) per month for focal stocks with the *lowest R*²'s, it is 0.267% (*t*-stat 2.581) with the *highest R*²'s, and the *difference* of 1.101% (*t*-stat 4.132) is highly statistically significant. This result is strong evidence of showing the key hypothesis of our theoretical discussions: assets with high SAR have high SARP, and assets with low SAR have low SARP. Beyond the aforementioned cross-sectional results, we also document the presence of an “unconditional SARP”. The “Avg” portfolio in Table 3 takes the simple average of returns of all of the ten $k = \text{‘Lo’}, 2, \dots, \text{‘Hi’}$ bins. We find that “Avg” enjoys a monthly SARP of 0.499% (*t*-stat 3.391).

In all, these empirical results show strong evidence in support of our core theoretical predictions: *(i)* *an unconditional SARP exists and is positive for the average representative asset; and (ii) SARP is increasing with SAR: stocks with low (high) SAR earn a low (high) SARP in the cross-section.*

3.3 Controlling for other risk factors and characteristics

We present two sets of evidence to show our main result — SARP is increasing in SAR — from Table 3 is robust after controlling for risk factors and other characteristics.

3.3.1 Risk adjusted returns on post-formation portfolios

We show our results still persist even after adjusting for the Fama-French three factors (Fama and French (1992, 1993)) and Fama-French five factors (Fama and French (2015)). For each bin k

and each return $r_t^k \in \{\bar{R}_{\text{Foc},t}^k, \bar{R}_{\text{Rep},t}^k, \bar{R}_{\text{LS},t}^k\}$, we consider the following two time series factor model regressions:

$$r_t^k = \alpha^k + \beta_{\text{MktRF}}^k \text{MktRF}_t + \beta_{\text{SMB}}^k \text{SMB}_t + \beta_{\text{HML}}^k \text{HML}_t + \varepsilon_t^k, \quad (14a)$$

$$r_t^k = \alpha^k + \beta_{\text{MktRF}}^k \text{MktRF}_t + \beta_{\text{SMB}}^k \text{SMB}_t + \beta_{\text{HML}}^k \text{HML}_t + \beta_{\text{CMA}}^k \text{CMA}_t + \beta_{\text{RMW}}^k \text{RMW}_t + \varepsilon_t^k. \quad (14b)$$

The regressors are well-known: “MktRF” is the market factor, “SMB” is the size factor, “HML” is the value factor, “CMA” is the investment factor, and “RMW” is the profitability factor.

Table 4 shows the FF3 and FF5 α estimates. Even controlling for known factors, the overall result that SARP is increasing in SAR remains robust. In particular, the FF3 α of SARP of stocks with the highest SAR is 0.573% (t -stat 4.216) per month, while it is 0.020% (t -stat 0.207) for stocks with the lowest SAR, and the cross-sectional difference is statistically significant at 0.553% (t -stat 3.039). The cross-sectional α estimate of SARP using the FF5 model is analogous.

3.4 Bivariate dependent sorts

We show our main result remains robust after we control for various stock characteristics. At the end of each month, stocks are first sorted by a characteristic into quintile portfolios. Then for a given characteristic and within each of its five portfolios, we further sort stocks based on the elastic-net R^2 into deciles bins. Each of these ten elastic-net R^2 bins are then averaged over their respective five characteristic portfolios. In all, the ten resulting elastic-net R^2 bins represent returns that control for a particular characteristic. All portfolios are equal-weighted.

Let’s describe the characteristics. $MktCap$ is market capitalization and B/M is book-to-market, and their quintiles are calculated using the NYSE breakpoints. $IdioVol$ is Ang et al. (2006)’s idiosyncratic volatility; the idiosyncratic volatility of stock i at month $t-1$ is the standard deviation of the residuals, arising from the OLS regression of past one year’s daily returns leading up until month $t-1$ onto the Fama-French three factors. $TotalVol$ is total volatility; it is the sample standard deviation of a stock i ’s past twelve months of daily returns up until month $t-1$. $IdioSkew$ is idiosyncratic skewness; it is calculated as the sample Pearson’s moment of coefficient skewness of

Table 4: Fama-French 3 and 5 factor regressions on portfolios sorted by elastic-net R². We run the Fama-French three (14a) and five (14b) factor time series regressions onto the elastic-net R² sorted decile portfolios. Stocks are sorted into deciles based on their elastic-net R² from the lowest (quantile 1, labelled “Lo”) to highest (quantile 10, labelled “Hi”). The column labelled “Lo - Hi” is the monthly return difference between the “Lo” bin and the “Hi” bin. The row labelled “Avg” is the simple average of the monthly excess returns across the ten $k = \text{‘Lo’}, 2, \dots, \text{‘Hi’}$ bins. The group labelled “Focal” reports the one-month ahead portfolio excess returns (7). The group labelled “Replicate” reports the one-month ahead excess returns of a portfolio that are constructed out of the estimated normalized elastic-net beta coefficients according to (9). The group labelled “Foc - Rep” reports the one-month ahead returns of the portfolio of a long position in the focal stocks, and a short position in the corresponding replicates. We define the *Statistical Arbitrage Risk Premium* (SARP) as the returns from “Foc - Rep”. In addition, a stock with low EN R² is said to have high *Statistical Arbitrage Risk* (SAR), while a stock with high EN R² is said to have low SAR. The estimated values are reported in monthly percentage points (e.g. 1.0 means 1%) and we report the Newey and West (1987) *t*-statistics with six lags in parentheses. The sample period is from January 31, 1976 to December 31, 2020.

	Lo EN R ²	2	3	4	5	6	7	8	9	Hi EN R ²	Lo - Hi
Focal											
FF3 α	0.658 (5.338)	0.360 (2.909)	0.155 (1.166)	0.061 (0.485)	0.000 (0.001)	-0.063 (-0.722)	-0.097 (-1.500)	-0.100 (-1.581)	-0.180 (-2.437)	-0.219 (-2.376)	0.877 (4.908)
FF5 α	0.700 (5.684)	0.383 (2.911)	0.140 (1.012)	0.044 (0.318)	-0.014 (-0.111)	-0.051 (-0.453)	-0.028 (-0.265)	0.060 (0.529)	0.061 (0.485)	-0.033 (-0.336)	0.733 (4.217)
Replicate											
FF3 α	0.087 (1.526)	0.198 (1.859)	-0.031 (-0.143)	-0.200 (-0.829)	-0.396 (-1.676)	-0.353 (-1.629)	-0.150 (-0.643)	-0.061 (-0.253)	-0.169 (-0.907)	-0.239 (-2.021)	0.326 (2.962)
FF5 α	0.100 (1.631)	0.195 (1.829)	-0.052 (-0.253)	-0.144 (-0.557)	-0.209 (-0.750)	-0.069 (-0.275)	0.172 (0.643)	0.297 (1.081)	0.160 (0.758)	-0.062 (-0.507)	0.162 (1.399)
Foc - Rep											
FF3 α	0.573 (4.216)	0.163 (1.142)	0.197 (1.040)	0.264 (1.271)	0.403 (1.913)	0.299 (1.437)	0.060 (0.275)	-0.045 (-0.209)	-0.008 (-0.055)	0.020 (0.207)	0.553 (3.039)
FF5 α	0.602 (4.330)	0.187 (1.231)	0.199 (1.041)	0.193 (0.900)	0.203 (0.909)	0.031 (0.148)	-0.193 (-0.854)	-0.242 (-1.081)	-0.098 (-0.618)	0.028 (0.281)	0.574 (3.149)

the residuals of the OLS regression of a stock's past twelve months' daily returns up until month $t-1$ onto the Fama-French three factors (with intercept). *TotalSkew* is total skewness; it is calculated as the sample Pearson's moment of coefficient skewness of a stock's past twelve months' daily returns up until month $t-1$. *AmihudIlliq* is Amihud (2002)'s illiquidity measure; it is defined as $\text{AmihudIlliq}_{i,t-1} := \frac{1}{D_{i,t-1}} \sum_{d=1}^{D_{i,t-1}} \frac{|R_{i,d}|}{\text{VOLD}_{i,d}}$, where $D_{i,t-1}$ is the total number of trading days of stock i in the past twelve months leading up to month $t-1$ and *VOLD* was as defined in Section 3.1.1. *Mom* is momentum; we define momentum of stock i at the end of month $t-1$ (i.e. the end of the estimation period) as the return of the stock during the 11-month period covering months $t-12$ through $t-2$. *STR* is Jegadeesh (1990) and Lehmann (1990)'s short-term reversal; it is defined as $\text{STR}_{i,t-1} = 100 \times R_{i,t-1}$.

Table 5 shows the results. We omit showing the results on the replicates for brevity. Overall, these results that control for the aforementioned characteristics are consistent with the unconditional results of Table 3. We pay special attention our results after controlling for size, value, *VOLD*, idiosyncratic volatility and total volatility. In particular, even though Table 1 shows a size, value and *VOLD* tilt for the elastic-net R^2 sorted stocks, the results here show that SARP is still persistent after we control for these characteristics.

3.5 Statistical Arbitrage Risk factor (SAR factor)

The main empirical objective of this paper is to show assets' SARP increases with SAR. However as hinted in the main results Section 3.2, conventional empirical asset pricing results would suggest that SAR is a priced factor. We can show the following corollary factor result. Let us define the *Statistical Arbitrage Risk factor (SAR factor)* with returns at month t as

$$R_{\text{SAR},t} := \bar{R}_{\text{LS},t}^{\text{Lo}} - \bar{R}_{\text{LS},t}^{\text{Hi}}. \quad (15)$$

where recall the definition of $\bar{R}_{\text{LS},t}^k$ in (13). In other words, the SAR factor is simply the difference in the SARP of the high SAR and low SAR stocks.

Let's investigate the price of risk λ_{SAR} of our SAR factor by the classical Fama and MacBeth

Table 5: Dependent bivariate sorts by characteristics and elastic-net R². This table reports mean returns in monthly percentage points (e.g. 1.0 means 1%) and Newey and West (1987) robust *t*-statistics with 6 lags in parentheses. We perform a bivariate dependent sort. Each month, we first sort stocks based on a characteristic (i.e. size, book-to-market, idiosyncratic volatility, total volatility, Amihud's illiquidity, momentum, short-term reversal, dollar volume liquidity, idiosyncratic skewness, and total skewness; see Section 3.4 for details and references on these characteristics) into quintiles portfolios. For a given characteristic and within each of its five portfolios, we sort stocks based on elastic-net R² into decile bins, from the lowest (quantile 1, labelled “Lo”) to highest (quantile 10, labelled “Hi”). The column labelled “Lo - Hi” is the monthly return difference between the “Lo” bin and the “Hi” bin. These ten elastic-net R² bins are then averaged over each of the five characteristic portfolios. Thus these ten elastic-net R² bins represent returns that control for a particular characteristic. All portfolios are equal-weighted. The sample period is monthly from January 31, 1976 to December 31, 2020.

		Lo EN R ²	2	3	4	5	6	7	8	9	Hi EN R ²	Lo - Hi
MktCap	Focal	1.389 (5.527)	1.101 (4.335)	0.995 (3.829)	0.937 (3.521)	0.811 (3.018)	0.826 (2.974)	0.880 (3.078)	0.954 (3.162)	0.911 (2.997)	0.921 (2.917)	0.468 (3.070)
	Foc - Rep	1.085 (4.863)	0.680 (3.207)	0.616 (3.380)	0.642 (3.691)	0.420 (2.255)	0.481 (2.726)	0.236 (1.168)	0.180 (0.989)	0.205 (1.438)	0.404 (3.600)	0.681 (3.737)
	Focal	1.434 (5.074)	1.150 (4.074)	0.954 (3.250)	1.002 (3.586)	0.931 (3.346)	0.920 (3.324)	0.911 (3.347)	0.936 (3.422)	0.926 (3.461)	0.882 (3.404)	0.552 (3.835)
	Foc - Rep	1.293 (4.716)	0.820 (3.268)	0.577 (2.446)	0.608 (3.009)	0.524 (2.556)	0.461 (2.405)	0.198 (0.972)	0.135 (0.660)	0.242 (1.524)	0.353 (3.395)	0.941 (3.516)
IdioVol	Focal	1.416 (5.453)	1.090 (4.285)	1.015 (3.851)	0.952 (3.549)	0.985 (3.651)	0.855 (3.040)	0.945 (3.234)	0.923 (3.116)	0.827 (2.749)	0.781 (2.451)	0.635 (4.031)
	Foc - Rep	1.196 (4.830)	0.641 (2.659)	0.665 (3.214)	0.610 (2.994)	0.663 (3.401)	0.402 (2.163)	0.241 (1.205)	0.159 (0.836)	0.159 (1.099)	0.300 (2.618)	0.896 (4.371)
	Focal	1.422 (5.244)	1.124 (4.252)	1.004 (3.642)	0.977 (3.584)	0.960 (3.463)	0.859 (3.043)	0.921 (3.216)	0.893 (3.091)	0.849 (2.921)	0.777 (2.657)	0.645 (4.557)
	Foc - Rep	1.276 (4.888)	0.676 (2.772)	0.639 (2.870)	0.633 (3.148)	0.604 (3.103)	0.436 (2.379)	0.160 (0.788)	0.141 (0.743)	0.174 (1.194)	0.295 (2.851)	0.981 (4.301)
AmihudIlliq	Focal	1.367 (5.436)	1.103 (4.261)	0.973 (3.747)	0.992 (3.653)	0.866 (3.175)	0.850 (3.116)	0.950 (3.362)	0.871 (2.970)	0.979 (3.390)	0.932 (3.258)	0.434 (3.071)
	Foc - Rep	0.996 (4.546)	0.730 (3.655)	0.583 (3.142)	0.724 (4.076)	0.510 (2.826)	0.473 (2.697)	0.300 (1.478)	0.099 (0.508)	0.295 (2.042)	0.419 (3.906)	0.577 (3.197)
	Focal	1.428 (5.151)	1.114 (3.901)	1.103 (3.856)	0.976 (3.429)	0.975 (3.470)	0.882 (3.167)	0.894 (3.267)	0.895 (3.288)	0.775 (2.779)	0.703 (2.626)	0.725 (5.282)
	Foc - Rep	1.311 (4.803)	0.739 (2.841)	0.725 (3.048)	0.561 (2.611)	0.611 (2.914)	0.432 (2.301)	0.155 (0.763)	0.110 (0.537)	0.092 (0.604)	0.216 (2.168)	1.095 (4.309)
STR	Focal	1.461 (5.252)	1.171 (4.069)	0.983 (3.464)	0.945 (3.199)	0.881 (3.146)	0.833 (3.038)	0.888 (3.187)	0.901 (3.314)	0.834 (3.010)	0.822 (3.027)	0.638 (4.213)
	Foc - Rep	1.332 (4.893)	0.825 (3.235)	0.616 (2.525)	0.528 (2.408)	0.526 (2.631)	0.334 (1.724)	0.162 (0.792)	0.095 (0.469)	0.195 (1.377)	0.320 (3.160)	1.012 (3.972)
	Focal	1.341 (5.153)	1.103 (4.142)	0.985 (3.700)	0.936 (3.427)	0.871 (3.153)	0.831 (3.020)	0.951 (3.337)	0.929 (3.246)	0.973 (3.491)	0.949 (3.511)	0.392 (2.782)
	Foc - Rep	1.033 (4.711)	0.745 (3.686)	0.610 (3.261)	0.603 (3.487)	0.528 (3.006)	0.459 (2.673)	0.247 (1.234)	0.170 (0.842)	0.303 (2.017)	0.423 (4.029)	0.611 (3.211)
VOLD	Focal	1.451 (5.408)	1.148 (4.159)	1.041 (3.494)	0.943 (3.180)	0.903 (3.218)	0.856 (3.098)	0.843 (2.983)	0.898 (3.195)	0.864 (3.035)	0.787 (2.885)	0.664 (4.310)
	Foc - Rep	1.300 (4.938)	0.773 (2.915)	0.711 (3.234)	0.538 (2.558)	0.563 (2.986)	0.375 (1.928)	0.202 (1.057)	0.061 (0.293)	0.169 (1.172)	0.287 (2.773)	1.013 (4.095)
	Focal	1.437 (5.438)	1.156 (4.175)	1.032 (3.478)	0.982 (3.397)	0.950 (3.348)	0.804 (2.926)	0.842 (2.973)	0.908 (3.207)	0.831 (2.898)	0.794 (2.897)	0.643 (4.267)
	Foc - Rep	1.271 (4.982)	0.758 (2.894)	0.685 (3.107)	0.613 (3.029)	0.596 (3.214)	0.342 (1.774)	0.175 (0.851)	0.064 (0.325)	0.185 (1.296)	0.290 (2.908)	0.980 (4.118)

(1973) regressions. We run the Fama and MacBeth (1973) with four different models: (i) SAR factor + CAPM; (ii) SAR factor + FF3 + momentum; (iii) SAR factor + FF5 + momentum; and (iv) FF5 + momentum as a benchmark. The data of the momentum factor is available from Kenneth French's website. Table 6(a) shows the correlations of our SAR factor along with the other asset pricing factors. Next is the choice of test assets. We first evaluate the price of risk of our SAR factor on the classical 25 portfolios formed on size and book-to-market (5×5). In addition, given our projection and replicate construction procedure of Section 2 is explicitly dependent on past returns, it is prudent to evaluate our SAR factor on test assets that are also sorted along a past return dependent characteristic. To this end, we will also consider the 25 portfolios formed on size and momentum (5×5) and the 25 portfolios formed on size and residual variance (5×5). Finally, we also evaluate our SAR factor on portfolios that are further sorted along corporate fundamentals, and consider the 25 portfolios formed on book-to-market and investments (5×5), and the 36 portfolios formed on size, operating profitability and investments ($2 \times 4 \times 4$). All test portfolios are equal-weighted and are obtained from French's website. Table 6 shows the main result of this section and the focus is the value of λ_{SAR} . The SAR factor has a positive price of risk, is statistically significant, and is robust across the various model specifications and test assets. Focusing on the SAR factor + FF5 + momentum model, we see the price of risk estimate for the SAR factor range from 0.990% to 1.631% per month, depending on the test asset.

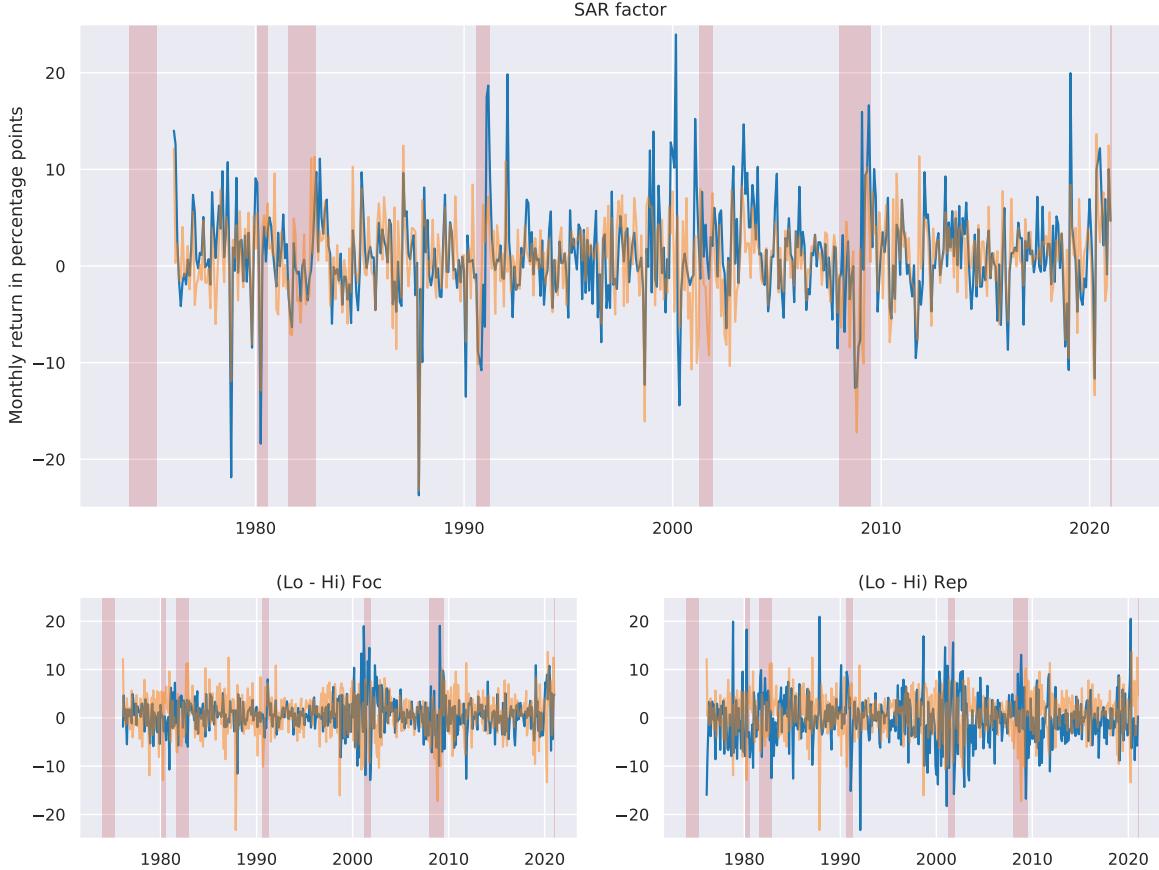
3.5.1 Investment strategy

The SAR factor is clearly a tradable portfolio. Figure 7(a) shows the cumulative returns (not inflation adjusted) from December 31, 1975 to December 31, 2020 of an initial \$100 investment on our SAR factor and other factors, and Figure 7(b) plots the log cumulative returns. The strategy “(Lo - Hi) Foc” is the cumulative returns from $\bar{R}_{\text{Foc},t}^{\text{Lo}} - \bar{R}_{\text{Foc},t}^{\text{Hi}}$; the strategy “(Lo - Hi) Rep” is that of $\bar{R}_{\text{Rep},t}^{\text{Lo}} - \bar{R}_{\text{Rep},t}^{\text{Hi}}$; and “SAR factor” is that of (15). Following the “(Lo - Hi) Foc” strategy will result in a \$3,177.63 cumulative return on December 31, 2020, “(Lo - Hi) Rep” will result in \$5.70 and SAR factor will result in \$18,219.66. The amazing performance of SAR factor is firstly because the “(Lo - Hi) Foc” strategy itself already has a high return. More importantly, however,

Table 6: SAR factor price of risk. This table shows the market price of risk λ_{SAR} in monthly percentage points (i.e. 1.0 means 1%) of the SAR factor and various other factors, and where the parentheses show the Fama and MacBeth (1973) t -statistics. The last column shows the χ^2 statistic and the brackets show its p -value. We define our *Statistical Arbitrage Risk factor* (SAR factor) in (15); the factor return $R_{SAR,t}$ is a result of a long position on the Lo elastic-net R^2 bin with returns $\bar{R}_{LS,t}^{Lo}$ and a short position on the Hi elastic R^2 bin with returns $\bar{R}_{LS,t}^{Hi}$. We focus on five equal-weighted tests assets: the 25 portfolios formed by the 5 size and 5 book-to-market bins; the 25 portfolios formed by the 5 size and 5 momentum bins; the 25 portfolios formed by size and Ang et al. (2006) residual variance bins; 25 portfolios formed by 5 book-to-market and 5 investment bins; and 36 portfolios formed by 2 size, 4 operating profitability and 4 investment bins. We use the Fama and MacBeth (1973) two-pass procedure by first running a monthly time series regression of the test assets onto the proposed factor models to obtain the betas, and then run a cross-sectional regression to obtain the prices of risk. The sampling period is monthly from January 31, 1976 to December 31, 2020.

Test asset	λ_{SAR}	λ_{MktRF}	λ_{SMB}	λ_{HML}	λ_{RMW}	λ_{CMA}	λ_{MOM}	χ^2
Size and B/M (5 × 5)								
	0.993 (4.695)	0.445 (3.308)	0.215 (1.575)	-0.171 (-0.724)	0.326 (1.861)	2.286 (4.122)	96.91 [p = 0.000]	
	0.652 (2.354)	0.777 (3.817)					64.66 [p = 0.000]	
	1.099 (3.989)	0.997 (4.774)	0.438 (3.267)	0.215 (1.589)		2.489 (4.449)	56.24 [p = 0.000]	
	1.224 (4.288)	0.861 (4.120)	0.434 (3.234)	0.184 (1.348)	0.366 (1.821)	-0.037 (-0.196)	1.876 (3.577)	161.15 [p = 0.000]
Size and Momentum (5 × 5)								
	0.717 (3.618)	0.308 (2.203)	1.053 (3.397)	-0.467 (-1.763)	0.894 (2.907)	0.615 (3.209)	64.50 [p = 0.000]	
	0.948 (3.492)	0.666 (3.291)					88.03 [p = 0.000]	
	1.097 (4.010)	0.674 (3.403)	0.334 (2.486)	0.782 (3.707)		0.616 (3.215)	59.11 [p = 0.000]	
	1.031 (3.633)	0.688 (3.462)	0.306 (2.192)	0.972 (3.121)	-0.206 (-0.738)	0.461 (1.370)	0.618 (3.223)	229.47 [p = 0.000]
Size and Residual Variance (5 × 5)								
	0.555 (2.809)	0.257 (1.705)	1.216 (5.292)	0.052 (0.340)	0.295 (0.968)	-0.541 (-0.822)	39.07 [p = 0.027]	
	0.406 (1.462)	0.832 (4.134)					581.88 [p = 0.000]	
	1.898 (4.998)	0.710 (3.595)	0.618 (3.967)	0.775 (3.601)		3.380 (4.358)	84.35 [p = 0.000]	
	1.631 (4.507)	0.678 (3.450)	0.434 (2.835)	1.037 (4.500)	0.225 (1.443)	-0.453 (-1.348)	2.379 (3.462)	74.61 [p = 0.000]
B/M and Investment (5 × 5)								
	0.921 (3.121)	0.177 (0.533)	0.638 (3.372)	-0.202 (-0.875)	0.826 (7.591)	0.751 (1.496)	11.53 [p = 0.985]	
	1.021 (3.666)	0.681 (3.240)					72.78 [p = 0.000]	
	1.868 (5.428)	1.190 (4.846)	-0.083 (-0.328)	0.490 (3.084)		2.615 (4.581)	84.89 [p = 0.000]	
	1.000 (2.573)	0.960 (3.262)	0.106 (0.326)	0.607 (3.088)	-0.116 (-0.458)	0.807 (7.129)	0.892 (1.567)	18.96 [p = 0.754]
Size, Operating profitability and Investment (2 × 4 × 4)								
	0.820 (3.998)	0.425 (2.726)	0.534 (2.677)	-0.007 (-0.061)	0.525 (5.034)	1.317 (3.020)	159.07 [p = 0.000]	
	0.766 (2.772)	0.836 (4.143)					223.96 [p = 0.000]	
	1.169 (4.145)	0.786 (3.801)	0.236 (1.556)	0.879 (5.312)		1.106 (2.287)	491.33 [p = 0.000]	
	0.990 (3.376)	0.808 (3.943)	0.339 (2.077)	0.668 (3.093)	0.019 (0.166)	0.477 (4.445)	1.232 (2.814)	136.89 [p = 0.000]

Figure 4: Time series plot of monthly returns of the SAR factor. The top panel plots the monthly returns of the SAR factor of (15) in blue, while the orange line is that of the market factor for comparison. Returns are expressed in monthly percentage points (e.g. 1.0 means 1%). The bottom left panel plots the strategy “(Lo - Hi) Foc”, while the bottom right panel plots the strategy “(Lo - Hi) Rep”. The red shaded regions are NBER recession dates. The sampling period is monthly from January 31, 1976 to December 31, 2020.



the SAR factor strategy simultaneously takes advantage of both the low returns of “(Lo - Hi) Rep” as leverage and also its high correlations with the focal stocks. As a matter of comparison, investing into the risk-free asset would result in the cumulative return of \$687.29; the market factor returns \$2,505.58; the SMB factor \$264.59; the HML factor \$220.58; the CMA factor \$318.83; the RMW factor \$467.17; and the MOM factor \$1,547.36. See Figure 4 for a plot of the monthly returns of the SAR factor. See Figure 5 for a plot of the rolling cumulative returns of the SAR factor.

Table 7: Cumulative returns and various descriptive statistics of the SAR factor and other assets. The column ‘‘Cum. return’’ of Table (a) shows the cumulative returns from an initial \$100 investment on December 31, 1975 to December 31, 2020 (the final amount is not inflation adjusted). The column ‘‘Ann. Sharpe ratio’’ shows the average monthly excess returns of various strategies divided by its standard deviation and then multiplied by $\sqrt{12}$. The third to sixth columns of (a) show the first four moments of the monthly excess returns time series; the mean and standard deviation columns are expressed in percentage points (e.g. 1.0 means 1%). The seventh column shows the 5-th and 95-th percentile of the monthly returns in percentage points. The column ‘‘ARMA(1, ·) coef’’ shows the estimated coefficient a while the column ‘‘ARMA(·, 1) coef’’ shows the estimated coefficient b of the ARMA(1,1) model $r_t = c + ar_{t-1} + b\epsilon_{t-1} + \epsilon_t$, and the parentheses show the associated t -statistic. Figure (b) shows the log cumulative returns from said initial investment across various investment strategies. The red shaded regions are NBER recession dates. Table (c) shows the time series correlations of our SAR factor against various other factors. The sampling period is monthly from January 31, 1976 to December 31, 2020.

	Cum. return	Ann. Sharpe ratio	Mean	Std	Skewness	Kurtosis	(5, 95)-th pct	ARMA(1, ·) coef	ARMA(·, 1) coef
(Lo - Hi) Foc	3,177.63	0.329	0.71	3.699	0.315	3.523	(-5.325, 6.239)	-0.95 (-11.87)	0.93 (10.53)
(Lo - Hi) Rep	5.70	-0.489	-0.387	5.308	0.053	2.258	(-8.898, 7.904)	-0.09 (-0.29)	0.20 (0.68)
SAR factor	18,219.66	0.495	1.101	5.169	0.07	3.456	(-6.293, 9.734)	0.36 (3.05)	-0.12 (-0.99)
Mkt-RF	2,505.58	0.263	0.7	4.477	-0.654	2.202	(-7.212, 7.191)	-0.68 (-3.24)	0.75 (3.92)
SMB	264.59	-0.160	0.223	2.914	0.207	3.83	(-4.111, 4.740)	0.34 (0.62)	-0.31 (-0.56)
HML	220.58	-0.199	0.19	2.948	0.079	2.463	(-4.082, 5.213)	0.58 (3.29)	-0.40 (-2.05)
RMW	467.17	-0.070	0.312	2.279	-0.4	12.862	(-2.583, 3.482)	0.07 (0.32)	0.07 (0.30)
CMA	318.83	-0.222	0.234	1.936	0.418	2.069	(-2.521, 3.284)	-0.58 (-2.87)	0.68 (3.80)
MOM	1,547.36	0.199	0.608	4.366	-1.335	10.738	(-6.555, 6.701)	-0.41 (-1.51)	0.49 (1.87)
RF	687.29		0.358	0.291	0.737	0.351	(0.000, 0.870)	0.98 (128.39)	-0.13 (-2.54)

(a) Cumulative returns, Sharpe ratios and moments

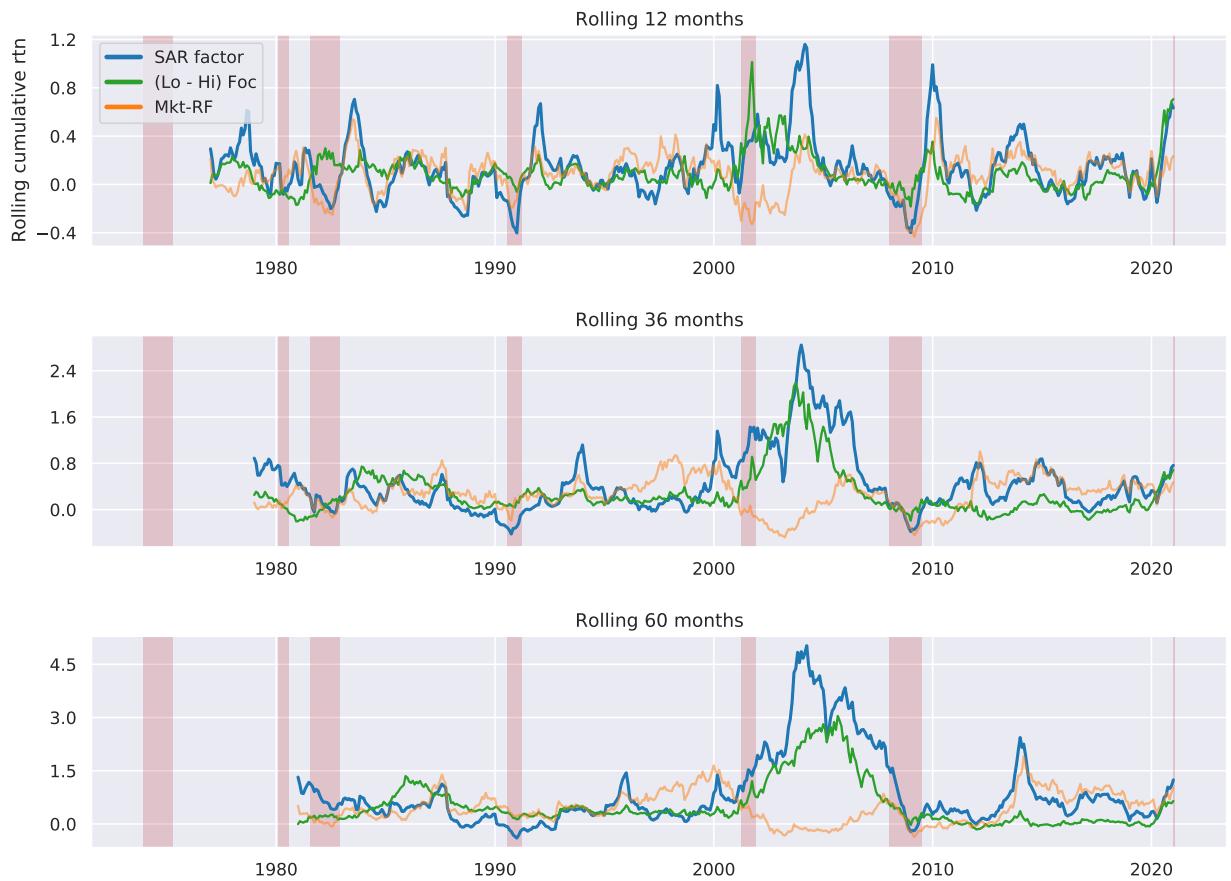


(b) Plot of log cumulative returns

	SAR	MktRF	SMB	HML	RMW	CMA	MOM	(Lo - Hi) Foc	(Lo - Hi) Rep
SAR	1.000								
MktRF	0.555	1.000							
SMB	0.593	0.257	1.000						
HML	-0.099	-0.209	-0.053	1.000					
RMW	-0.385	-0.259	-0.399	0.198	1.000				
CMA	-0.107	-0.359	-0.051	0.674	0.113	1.000			
MOM	-0.090	-0.133	0.002	-0.225	0.098	-0.018	1.000		
(Lo - Hi) Foc	0.315	-0.361	0.025	0.142	0.107	0.285	0.181	1.000	
(Lo - Hi) Rep	-0.752	-0.790	-0.559	0.193	0.448	0.301	0.213	0.389	1.000

(c) Correlations

Figure 5: Rolling window cumulative returns of the SAR factor. We plot the cumulative returns of rolling $n = 12, 36$ and 60 months of an initial \$1 investment of the SAR factor, the “(Lo - Hi) Foc” strategy, and the market factor for comparison (e.g. 1.5 means a total of return of 150% over the past n many months). The red shaded regions are NBER recession dates. The sampling period is monthly from January 31, 1976 to December 31, 2020.



3.6 Additional robustness checks

We run several robustness checks on our main result and they are summarized in Table 8.

3.6.1 SARP is not driven by the equity risk premium

Is SARP just the equity risk premium in disguise? From Table 1(iv) and the replicate construction procedure of Section 2.2, it is evident that stocks with high SAR have replicates that are essentially just the risk-free asset. The SARP for high SAR stocks is simply just the equity premium. However, the SARP for low SAR stocks are distinctively different from their equity premia; there are a plethora of risky peers in the replicates of low SAR stocks. Nonetheless, as constructed, there is potential concern the main empirical message of this paper — stocks with high (low) SAR have high (low) SARP — is driven by the equity premium of low SAR stocks. To rule out this concern, we follow the projection and replicate construction procedures of Section 2, but then drop all stocks whose replicates only consist of the risk-free assets (that is, the replicates now must contain at least one risky peer). Table 8(i) shows the result; compare this against the main result Table 3. We see high SAR stocks still have a monthly return of 0.658% (t -stat 3.901) higher than that of low SAR stocks. More importantly, we see the SARP of high SAR stocks (which is now distinctly different from the equity risk premium) is still substantially higher at 0.815% (t -stat 3.147) per month than the SARP of low SAR stocks. Hence the cross-sectional difference in SARP is not driven by the equity risk premium of high SAR stocks. In results (ii) to (vii), we repeat the bivariate sort procedure of Table 5 and sort various characteristics into quintiles. Our main result remains robust after subsetting for stocks to have at least one risky peer in its replicate, and after controlling for these characteristics.

3.6.2 Weak SARP when using FF3 and FF5 for replicate construction

Is there any value at all in running a computationally expensive elastic-net projection procedure of Section 2.1.2? Can we get analogous results using simple OLS projections with much fewer regressors? Instead of projecting each stock i onto the past twelve months' daily returns of every other stock using elastic-net, let's simply project each stock i 's past twelve months' daily returns

Table 8: Robustness checks. The reported numbers are monthly returns in percentage points (i.e. 1.0 means 1.0%) and the parentheses show Newey and West (1987) robust t -statistics with six lags. Result (i) repeats the procedure that shows the main result Table 3 except we restrict each stock to have a replicate that must contain at least one single risky asset, so the subsetted stocks' replicates cannot simply just be the risk-free asset. Results (ii) to (iv) repeat the bivariate dependent sort procedure of Table 5 but enforces the aforementioned minimum one risky asset requirement on the stocks. Results (viii) and (x) are analogous to the main result of Table 3 except we sort stocks by deciles by their FF3 and FF5 R^2 's, respectively. FF3 and FF5 replicates for each stock are constructed out of projected normalized OLS coefficients according to Section 3.6.2. Results (ix) and (xi) show the corresponding FF3 and FF5 R^2 decile sort results that first sort stocks into their market capitalization into quintiles; compare analogously to Table 5 for bivariate elastic-net R^2 sort results. The sample period is monthly from January 31, 1976 to December 31, 2020.

		Lo	2	3	4	5	6	7	8	9	Hi	$Lo - Hi$
(i) EN R^2 with min one risky asset	Focal	1.440 (5.051)	1.001 (3.444)	1.026 (3.407)	0.888 (3.128)	0.987 (3.514)	0.946 (3.425)	0.864 (3.104)	0.865 (3.143)	0.793 (2.817)	0.782 (2.958)	0.658 (3.901)
	Foc - Rep	1.118 (4.077)	0.617 (2.382)	0.560 (2.468)	0.410 (1.933)	0.253 (1.109)	0.155 (0.674)	0.043 (0.201)	0.134 (0.817)	0.252 (2.058)	0.303 (2.994)	0.815 (3.147)
(ii) ——— controlling for MktCap	Focal	1.332 (5.330)	1.019 (4.027)	0.893 (3.486)	0.837 (3.227)	0.930 (3.423)	0.886 (3.123)	0.942 (3.228)	0.950 (3.148)	0.898 (2.959)	0.894 (2.858)	0.438 (2.826)
	Foc - Rep	0.835 (3.589)	0.580 (2.960)	0.477 (2.488)	0.443 (2.194)	0.312 (1.420)	0.080 (0.366)	0.205 (1.064)	0.177 (1.080)	0.297 (2.657)	0.419 (3.744)	0.416 (2.163)
(iii) ——— controlling for B/M	Focal	1.380 (4.868)	0.988 (3.388)	0.998 (3.471)	0.952 (3.426)	0.981 (3.592)	0.912 (3.288)	0.916 (3.329)	0.931 (3.411)	0.912 (3.404)	0.888 (3.418)	0.492 (3.350)
	Foc - Rep	1.068 (4.110)	0.596 (2.484)	0.543 (2.467)	0.442 (2.155)	0.259 (1.188)	0.122 (0.518)	0.101 (0.494)	0.201 (1.139)	0.289 (2.218)	0.396 (3.831)	0.672 (2.729)
(iv) ——— controlling for IdioVol	Focal	1.354 (5.278)	1.003 (3.887)	1.013 (3.895)	0.929 (3.444)	0.950 (3.513)	0.933 (3.268)	0.958 (3.291)	0.869 (2.966)	0.859 (2.825)	0.759 (2.418)	0.595 (3.669)
	Foc - Rep	0.916 (3.593)	0.543 (2.446)	0.606 (2.745)	0.529 (2.603)	0.215 (0.965)	0.159 (0.736)	0.168 (0.832)	0.173 (1.071)	0.252 (2.133)	0.315 (2.797)	0.601 (2.672)
(v) ——— controlling for Mom	Focal	1.349 (4.794)	1.130 (3.955)	1.003 (3.464)	0.991 (3.540)	0.964 (3.463)	0.919 (3.333)	0.924 (3.384)	0.830 (3.042)	0.790 (2.809)	0.703 (2.628)	0.646 (4.617)
	Foc - Rep	1.008 (3.771)	0.736 (3.068)	0.593 (2.782)	0.488 (2.238)	0.201 (0.954)	0.132 (0.570)	0.139 (0.638)	0.087 (0.488)	0.228 (1.891)	0.228 (2.278)	0.780 (3.216)
(vi) ——— controlling for STR	Focal	1.435 (4.999)	1.008 (3.611)	0.925 (3.116)	0.934 (3.319)	0.954 (3.429)	0.849 (3.072)	0.949 (3.471)	0.879 (3.183)	0.830 (2.947)	0.809 (3.004)	0.626 (3.820)
	Foc - Rep	1.116 (4.087)	0.606 (2.490)	0.486 (2.168)	0.445 (2.146)	0.247 (1.121)	0.005 (0.023)	0.133 (0.628)	0.148 (0.920)	0.302 (2.555)	0.322 (3.209)	0.795 (3.156)
(vii) ——— controlling for VOLD	Focal	1.259 (4.905)	1.010 (3.807)	0.921 (3.362)	0.873 (3.296)	0.959 (3.520)	0.871 (3.097)	0.971 (3.367)	0.929 (3.228)	0.958 (3.491)	0.928 (3.443)	0.331 (2.329)
	Foc - Rep	0.800 (3.623)	0.544 (2.671)	0.534 (2.745)	0.507 (2.675)	0.223 (0.968)	0.132 (0.643)	0.197 (0.939)	0.200 (1.182)	0.383 (3.028)	0.410 (3.899)	0.390 (2.197)
(viii) FF3 OLS R^2	Focal	1.260 (3.961)	1.140 (3.667)	1.006 (3.245)	0.971 (3.131)	0.974 (3.363)	0.907 (3.197)	0.907 (3.421)	0.880 (3.497)	0.880 (3.482)	0.806 (3.165)	0.455 (1.944)
	Foc - Rep	1.121 (4.353)	0.821 (3.787)	0.630 (2.998)	0.487 (2.546)	0.445 (2.087)	0.386 (2.490)	0.416 (2.092)	0.364 (1.749)	0.319 (1.399)	0.245 (1.216)	0.193 (2.882)
(ix) ——— controlling for MktCap	Focal	0.872 (3.678)	0.961 (3.801)	0.977 (3.754)	1.002 (3.635)	1.026 (3.755)	0.994 (3.654)	0.997 (3.530)	1.015 (3.413)	0.953 (3.074)	0.923 (2.879)	-0.052 (-0.274)
	Foc - Rep	0.606 (3.645)	0.511 (3.046)	0.542 (3.115)	0.457 (2.713)	0.550 (3.291)	0.519 (3.606)	0.481 (3.004)	0.494 (3.336)	0.414 (2.778)	0.371 (2.620)	0.235 (1.672)
(x) FF5 OLS R^2	Focal	1.245 (4.029)	1.147 (3.612)	1.009 (3.237)	0.989 (3.213)	0.991 (3.396)	0.916 (3.246)	0.886 (3.350)	0.874 (3.482)	0.880 (3.509)	0.792 (3.097)	0.453 (1.998)
	Foc - Rep	1.093 (4.484)	0.868 (3.838)	0.626 (3.058)	0.460 (2.280)	0.433 (2.502)	0.404 (2.340)	0.335 (1.927)	0.313 (1.729)	0.255 (1.461)	0.191 (1.247)	0.902 (2.960)
(xi) ——— controlling for MktCap	Focal	0.881 (3.707)	0.925 (3.732)	1.048 (4.086)	1.010 (3.702)	1.023 (3.699)	0.981 (3.529)	0.953 (3.363)	1.019 (3.408)	0.962 (3.171)	0.916 (2.819)	-0.035 (-0.182)
	Foc - Rep	0.584 (3.516)	0.554 (3.294)	0.548 (3.190)	0.502 (2.884)	0.514 (3.137)	0.510 (3.311)	0.432 (2.678)	0.448 (3.064)	0.451 (3.160)	0.400 (2.877)	0.185 (1.360)

onto the FF3 and FF5 factors using OLS.⁷ For each stock i and at month $t - 1$, we collect the 3×1 FF3 OLS projection vector $\hat{\beta}_{i,t-1}^{\text{FF3}}$ and the 5×1 FF5 OLS projection vector $\hat{\beta}_{i,t-1}^{\text{FF5}}$. Symmetric to the normalizing procedure of Section 2.2, we will normalize the projection vectors by their L^1 norm as $\tilde{\beta}_{i,t-1}^{\text{FF}z} := \hat{\beta}_{i,t-1}^{\text{FF}z} / ||\hat{\beta}_{i,t-1}^{\text{FF}z}||_1$ for $z = 3, 5$. Analogous to (9), the returns of the $\text{FF}z$ replicate of stock i is

$$R_{i,t}^{\text{Rep, FF}z} := (1 - \mathbf{1}^\top \tilde{\beta}_{i,t-1}^{\text{FF}z}) r_{f,t} + \mathbf{R}_{\text{FF}z,t}^\top \tilde{\beta}_{i,t-1}^{\text{FF}z}, \quad (16)$$

where $\mathbf{R}_{\text{FF}3,t} = [R_{\text{MktRF},t}, R_{\text{SMB},t}, R_{\text{HML},t}]^\top$ and $\mathbf{R}_{\text{FF}5,t} = [R_{\text{MktRF},t}, R_{\text{SMB},t}, R_{\text{HML},t}, R_{\text{CMA},t}, R_{\text{RMW},t}]^\top$ are the month t returns of the FF3 and FF5 factors. Finally, we now sort all stocks on their $\text{FF}z$ OLS R^2 's into deciles.

Table 8(viii) shows the FF3 results while (x) shows the FF5 results. At first glance, it seems we can still identify SARP in the cross-section of $\text{FF}z$ R^2 decile sorted stocks. The FF3 procedure shows that high FF3 SAR stocks have a SARP of 1.121% (t -stat 4.353) per month while low FF3 SAR stocks have a 0.193% (t -stat 1.216), with a cross-sectional difference of 0.928% (t -stat 2.882). We can see a similar result in the FF5 projections in (iv). However upon closer inspection, the $\text{FF}z$ procedures are not robust to size effects, while it is robust for the elastic-net counterpart of Section 2.1.2. In Table 8(ix), we repeat the bivariate sort procedure to that of Table 5 but with FF3 R^2 decile bins. After controlling for stocks' market capitalization quintiles, there is no cross-sectional SARP: stocks with high FF3 SAR have a SARP of 0.606% (t -stat 3.645) while stocks with low FF3 SAR have a SARP of 0.371% (t -stat 2.620), but the cross-sectional difference of 0.235% (t -stat 1.672) is statistically insignificant. Likewise in result (xi), we see no cross-sectional SARP in stocks that are constructed out of the aforementioned FF5 projection and sorting procedure, after controlling for size. Overall, these results show SARP can still be found using conventional asset pricing factor models and the standard OLS procedure. However, the “quality” of this SARP is rather low compared to the SARP that is identified and constructed via our elastic-net procedure.

⁷Since we did not project onto an intercept using the elastic-net estimator, we will also not project an intercept when using the FF3 and FF5 OLS regressions (14a) and (14b). Recall Remark 2.2.

4 Conclusion

This paper theoretically and empirically show the relationship between the *Statistical Arbitrage Risk* (SAR) and *Statistical Arbitrage Risk Premium* (SARP) of a stock. Theoretically, SARP is the expected return of the residual factor risks of a given stock. Empirically we use the elastic-net, a machine learning method, to project a given stock’s returns onto the span of every other stock in the market. The projection R^2 is SAR, and the normalized projection coefficient entries serve as investment weights in constructing the replicate portfolio of a given stock. The core message of this paper is: *SARP is increasing in SAR*.

We see several interesting directions to further study SAR and SARP by machine learning (ML). In this paper, the elastic-net serves two simultaneous roles and steps: (1) *fitting and variable selection* and (2) *portfolio construction*. In Step (1), the elastic-net R^2 is used to measure the SAR of a stock i , and the non-zero entries of the projection vector are used to identify the risky peers of stock i . In Step (2), the normalized values of the non-entries are used as investment weights to construct the replicate of stock i . A further study of SAR and SARP would be to disentangle these two roles by using two different ML methods. Step (1) can benefit from using a more sophisticated method ML_1 that take in “wide regressors” as inputs but has better sparse variable selection and goodness-of-fit R^2 (and hence SAR) properties than the elastic-net. For step (2), we see benefits in designing a ML_2 replicate portfolio construction method that also takes into account the relative importance of these peers from step (1). Once these two improved steps are complete, the estimation of SARP and other analyses can follow conventional empirical asset pricing procedures as outlined in this paper. Tangential to using ML methods for forecasting and factors selection in finance, we feel a novel avenue for ML applications in finance is this sense of portfolio construction. In closing, we strongly believe this paper has merely scratched the surface in the study of SAR and SARP by machine learning.

References

- AMIHUD, Y. (2002): “Illiquidity and stock returns: cross-section and time-series effects,” *Journal of Financial Markets*, 5, 31–56.
- ANG, A., R. J. HODRICK, Y. XING, AND X. ZHANG (2006): “The Cross-Section of Volatility and Expected Returns,” *Journal of Finance*, 61, 259–299.
- AVELLANEDA, M. AND J.-H. LEE (2010): “Statistical arbitrage in the US equities market,” *Quantitative Finance*, 10, 761–782.
- BALI, T. G., R. F. ENGLE, AND S. MURRAY (2016): *Empirical Asset Pricing: The Cross Section of Stock Returns*, John Wiley & Sons.
- BLACK, F., M. C. JENSEN, AND M. SCHOLES (1972): “The capital asset pricing model: Some empirical tests,” *Studies in the theory of capital markets*, 81, 79–121.
- BREEDEN, D. T., M. R. GIBBONS, AND R. H. LITZENBERGER (1989): “Empirical tests of the consumption-orientated CAPM,” *Journal of Finance*, 44, 231–262.
- CHINCO, A., A. D. CLARK-JOSEPH, AND M. YE (2019): “Sparse signals in the cross-section of returns,” *The Journal of Finance*, 74, 449–492.
- FAMA, E. F. AND K. R. FRENCH (1992): “The cross-section of expected stock returns,” *Journal of Finance*, 47, 427–465.
- (1993): “Common risk factors in the returns on stocks and bonds,” *Journal of Financial Economics*, 33, 3 – 56.
- (2015): “A five-factor asset pricing model,” *Journal of Financial Economics*, 116, 1 – 22.
- FAMA, E. F. AND J. D. MACBETH (1973): “Risk, return, and equilibrium: Empirical tests,” *Journal of political economy*, 81, 607–636.
- FENG, G., S. GIGLIO, AND D. XIU (2020): “Taming the factor zoo: A test of new factors,” *The Journal of Finance*, 75, 1327–1370.

- FREYBERGER, J., A. NEUHIERL, AND M. WEBER (2017): “Dissecting Characteristics Nonparametrically,” Working Paper 23227, National Bureau of Economic Research.
- GATEV, E., W. N. GOETZMANN, AND K. G. ROUWENHORST (2006): “Pairs trading: Performance of a relative-value arbitrage rule,” *The Review of Financial Studies*, 19, 797–827.
- GU, S., B. KELLY, AND D. XIU (2020): “Empirical asset pricing via machine learning,” *The Review of Financial Studies*, 33, 2223–2273.
- HANSEN, L. P. AND R. J. HODRICK (1980): “Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis,” *Journal of Political Economy*, 88, 829–853.
- HARVEY, C. R., Y. LIU, AND H. ZHU (2016): “... and the Cross-Section of Expected Returns,” *Review of Financial Studies*, 29, 5–68.
- HEDEGAARD, E. AND R. J. HODRICK (2016): “Estimating the risk-return trade-off with overlapping data inference,” *Journal of Banking and Finance*, 67, 135 – 145.
- HUCK, N. (2019): “Large data sets and machine learning: Applications to statistical arbitrage,” *European Journal of Operational Research*, 278, 330–342.
- JEGADEESH, N. (1990): “Evidence of predictable behavior of security returns,” *Journal of Finance*, 45, 881–898.
- KRAUSS, C. (2017): “Statistical arbitrage pairs trading strategies: Review and outlook,” *Journal of Economic Surveys*, 31, 513–545.
- LAMONT, O. A. (2001): “Economic tracking portfolios,” *Journal of Econometrics*, 105, 161–184.
- LEHMANN, B. N. (1990): “Fads, martingales, and market efficiency,” *Quarterly Journal of Economics*, 105, 1–28.
- LEUNG, R. C. W. AND Y.-M. TAM (2021): “Online Supplementary Materials for ‘Statistical Arbitrage Risk Premium by Machine Learning’,” Tech. rep.

- LINTNER, J. (1965): “The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets,” *The Review of Economics and Statistics*, 47, 13–37.
- NEWHEY, W. K. AND K. D. WEST (1987): “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix,” *Econometrica*, 703–708.
- Ross, S. (1976): “The Arbitrage Theory of Capital Asset Pricing,” *Journal of Economic Theory*, 13, 341–360.
- SHARPE, W. F. (1964): “Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk,” *Journal of Finance*, 19, 425–442.
- SHU, L., F. SHI, AND G. TIAN (2020): “High-dimensional index tracking based on the adaptive elastic net,” *Quantitative Finance*, 1–18.
- TIBSHIRANI, R. (1996): “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- WURGLER, J. AND E. ZHURAVSKAYA (2002): “Does arbitrage flatten demand curves for stocks?” *The Journal of Business*, 75, 583–608.
- ZOU, H. AND T. HASTIE (2005): “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

A Appendix

A.1 Discussions on the projection procedure

Here we outline the details of our projection procedure.

A.1.1 Elastic-net estimator

Let's introduce the optimization problem of the *elastic-net* estimator developed by Zou and Hastie (2005). Let y be a $T \times 1$ vector, \mathbf{X} be a $T \times N$ matrix and let β be a $N \times 1$ vector. Consider the optimization problem,

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^N} \left\{ \frac{1}{2T} \|y - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\}, \quad \lambda_1, \lambda_2 \geq 0, \quad (17)$$

where $\|\mathbf{x}\|_1 := \sum_{j=1}^N |x_j|$ is the L^1 -norm on \mathbb{R}^N , and $\|\mathbf{x}\|_2 := \sqrt{\sum_{j=1}^N x_j^2}$ is the L^2 -norm on \mathbb{R}^N . In our application, y will be a time series vector of T days of returns of a particular stock, and \mathbf{X} will be the concatenation of the time series of N number of other stocks. Note this means there are $N + 1$ total number of stocks.

The solution $\hat{\beta}$ is called the *elastic-net* estimator. This estimator encompasses the special cases of the *ordinary least squares (OLS)* estimator (when $\lambda_1 = 0, \lambda_2 = 0$), *least absolute shrinkage and selection operator (LASSO)* estimator of Tibshirani (1996) (when $\lambda_1 > 0, \lambda_2 = 0$), and the *ridge* estimator (when $\lambda_1 = 0, \lambda_2 > 0$). The hyperparameters λ_1, λ_2 control the strength of the L^1 - and L^2 -norm penalties, respectively. In this paper when we refer to the elastic-net estimator, we always refer to the case when λ_1, λ_2 are both strictly positive. In our actual implementation, we use a 3-fold *cross-validation* procedure to empirically select the hyperparameters $\lambda_1, \lambda_2 > 0$.

Remark A.1. To reduce the already lengthy computational time in estimating (17), we cross-validate for only one hyperparameter rather than two in our empirical studies by making the following simplifying assumption on λ_1, λ_2 . We set $\lambda_1 = \lambda\ell$ and $\lambda_2 = \frac{1}{2}\lambda(1 - \ell)$, and set $\ell = 1/2$, and only cross-validate for the single $\lambda > 0$ parameter.

A.1.2 Intercept estimation, stale prices and sparsity

We deliberately do *not* estimate an intercept in (17). ⁸ This is to avoid attributing a price with very stale prices with high R^2 .

Imagine a given stock i has a vector of 12 months' worth daily returns $y_{i,t-1}$, where all the daily returns are almost all 0's except on a handful of days. As an extreme, suppose all other stocks has a return matrix $\mathbf{X}_{i,t-1}$ of (6) with rank $D_{i,t-1}$. ⁹ Then using the elastic-net estimator without intercept (17), the squared error term would be high, and thus leading to an overall low $R^2_{i,t-1}$. Observe that in this case $\hat{\beta}_{i,t-1} = \mathbf{0}$ is not necessarily an optimal solution because $\mathbf{X}_{i,t-1}$ is of rank $D_{i,t-1}$ while $\hat{\beta}_{i,t-1}$ is $N_{t-1} \times 1$, and we have that $D_{i,t-1} \ll N_{t-1}$. This means there could exist some sparse $\hat{\beta}_{i,t-1} \neq \mathbf{0}$ that achieves a smaller value in (17) than that of the zero vector. This is so when $\|y_{i,t-1} - \mathbf{X}_{i,t-1}\hat{\beta}_{i,t-1}\|_2^2 + \lambda_1\|\hat{\beta}_{i,t-1}\|_1 + \lambda_2\|\hat{\beta}_{i,t-1}\|_2^2 \ll \|y_{i,t-1}\|_2^2$, especially when the penalty weights λ_1, λ_2 are small. In contrast, if one were to use the elastic-net estimator *with* an intercept, then setting $\hat{\beta}_{i,t-1} = \mathbf{0}$ with intercept $\hat{c} \approx 0$ is an optimal solution. As a result, this would lead to the numerator term in the calculation of R^2 to approximately equal to zero, and thus resulting in a high R^2 . We want to *avoid* the latter case in our results.

In this paper, we want to exclusively reserve “high R^2 ” to mean stock i can be well explained by other risky assets, and “low R^2 ” to mean stock i cannot be explained by other risky assets.

A.1.3 Parsimony: Why elastic-net and not other machine learning methods?

Out of a myriad of machine learning methods, why did we choose the elastic-net estimator to test our empirical implication? In Theorem 1.1, we motivated the need to *linearly* regress a stock return R_i onto the returns \mathbf{R}_{-i} of all other stocks. The elastic-net estimator can actually be seen as a linear regression problem with constraints. By Lagrange-duality, (17) is identical to the following

⁸The elastic-net estimator that contains an intercept is given by,

$$\hat{c}, \hat{\beta} \in \arg \min_{c \in \mathbb{R}, \beta \in \mathbb{R}^p} \left\{ \|y - c - \mathbf{X}\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \right\}, \quad \lambda_1, \lambda_2 \geq 0,$$

where by convention, L^1 - and L^2 -penalties are only applied on the β coefficients and not on the intercept c .

⁹In the actual empirical implementations, we do not impose nor check for any such condition.

constrained least squares problem,

$$\begin{aligned} \min_{\beta \in \mathbb{R}^N} \quad & \frac{1}{2T} \|y - \mathbf{X}\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq \eta_1, \\ & \|\beta\|_2^2 \leq \eta_2, \end{aligned} \tag{18}$$

for some $\eta_1, \eta_2 \geq 0$ that is dependent on the values of $\lambda_1, \lambda_2 \geq 0$. Most references in the statistical literature (e.g. Zou and Hastie (2005)) prefers the unconstrained Lagrangian-form (17) over the constrained optimization form (18) for theoretical and computational reasons. Here, we explicitly draw out the “linear nature” of the elastic-net estimator via (18) because Theorem 1.1 indeed predicts a linear relationship.

We can now justify why we choose to implement the elastic-net estimator in this paper out of the myriad of machine learning methods. Without a doubt, there are more advanced machine learning and econometric methods that can increase the in-sample fit and thereby boost the in-sample R^2 . A recent extensive study by Gu, Kelly, and Xiu (2020) shows that many possible machine learning methods can achieve high in-sample and out-of-sample R^2 ’s. However, the issue with these “black box” methods is that regardless of their in-sample or even out-of-sample performance, their estimated model parameters often do not have a transparent link to the regressors themselves. In contrast, although the elastic-net estimator is inherently non-linear, its estimated coefficients can be applied in linearly back to the regressors. This parsimonious nature of the elastic-net allows us to explicitly and linearly construct the replicates as in (9).

We emphasize that the elastic-net machine learning method is simply a tool to test the prediction of Theorem 1.1 via the construction of the replicates (9). We have no intentions to conduct statistical inference on the estimated elastic-net coefficients. Our statistical inference claims are still based upon well understood portfolio sort methods in the empirical asset pricing literature as outlined in Section 2.3.

The consideration of the replicates is what drives us to prefer the elastic-net over its close cousin, the LASSO. The LASSO enjoys a “sparsity property”, whereby the number of estimated coefficients

tend to be small, even though the set of regressors could be large. Sparsity means there are only a handful of stocks that have to be considered in order to construct the replicate of stock i . This is important because if the number of stocks needed to construct the replicate of stock i is large, then whatever statistical results we claim to find may be economically infeasible due to trading costs and other market frictions. Unfortunately, if a group of regressors are highly correlated with each other then LASSO has a tendency to only select one regressor effectively at random. This makes for a poor portfolio construction of the replicates for numerous reasons, and loss of potential diversification is an obvious one. The elastic-net remedies this problem by inheriting the grouping property of the ridge estimator. In all, this means the elastic-net is a good candidate for our consideration of the replicates because: (i) it can fit the data (i.e., from the least squares property of the OLS); (ii) it can linearly apply its estimated coefficients over the regressors; (iii) it has a sparsity property (i.e., from LASSO); and (iv) it has a grouping property (i.e., from ridge).

Would a more general machine learning method be useful for our purposes? The answer is mixed. A general form of a machine learning estimator has the form $y = g(\mathbf{X}; \theta, \lambda) + \varepsilon$, where y is the response variable, g could be a parameterized or non-parametric function, \mathbf{X} is the set of regressors, θ parameterizes g , λ is a hyperparameter, and ε is a nuisance parameter. Generally speaking, the relationship between \mathbf{X} , θ and λ could be highly non-linear. The data is typically split into three sets $[y_{\text{train}}, y_{\text{validate}}, y_{\text{test}}]$ and $[\mathbf{X}_{\text{train}}, \mathbf{X}_{\text{validate}}, \mathbf{X}_{\text{test}}]$. For a given hyperparameter λ , the machine learning method uses the *training set* ($y_{\text{train}}, \mathbf{X}_{\text{train}}$) to fit the data to find an optimal parameter $\hat{\theta}(\lambda)$. Using the *validation set* ($y_{\text{valid}}, \mathbf{X}_{\text{valid}}$), the method then finds an optimal $\hat{\lambda}$ that fits the validation data, and a trained model is then given by the parameters $\hat{\theta}(\hat{\lambda})$ and $\hat{\lambda}$. Finally, the forecast accuracy of the model is tested against the *testing set* ($y_{\text{test}}, \mathbf{X}_{\text{test}}$) by comparing against the predicted value $\hat{y} = g(\mathbf{X}_{\text{train}}; \hat{\theta}(\hat{\lambda}), \hat{\lambda})$ against its realization y_{train} .

In our context, if we were to apply such general machine learning method to asset i , then we would still use the same response variable and regressors as in (6). However, the fitted value would be of very little use to us when constructing the replicates (9) from a finance perspective. There are two issues. Firstly, we are *not* using a forecast value \hat{y} to construct portfolios. We are using the fitted coefficients $\hat{\theta}(\lambda) = \hat{\beta}_{i,t-1}$ (which is then subsequently normalized by its L^1 norm) as

investment weights for the replicate of stock i . In order to use fitted coefficients as investment weights, then it is almost *necessary* that these fitted coefficients have a clear and linear relationship to the one-month ahead returns $\mathbf{X}_{\text{train}} = \mathbf{R}_{-i,t}$. The OLS, LASSO and elastic-net certainly have this property, as seen in (9). However, a general machine learning method does *not* have this linear relationship between the fitted coefficients and its out-of-sample regressors. This non-linearity between fitted coefficients and its out-of-sample regressors explicitly prevent us from directly using these general machine learning methods in constructing a portfolio.

Nonetheless, as discussed at the conclusion of the main text, there are many research avenues to disentangle the fitting role and replicate portfolio construction role of the elastic-net. By designing a method ML_1 for fitting and another method ML_2 for replicate portfolio construction, there are many other properties of the SAR and SARP to explore. Regardless, and at least to us, seems to be the most parsimonious and is a good baseline.