Feature-Selection Risk*

Alex Chinco[†]

January 20, 2021

Abstract

Companies have overlapping exposures to many different features that might plausibly affect their returns, like whether they were involved in a crowded trade, whether they were mentioned in an M&A rumor, or whether their supplier recently missed an earnings forecast. At any point in time, only a handful of these features actually matter. Traders have to simultaneously infer both the identity and the value of the few relevant features. First, I show theoretically that, when traders face this sort of joint inference problem, the risk of selecting the wrong features can spill over and distort how they value assets. Then, I empirically document that using an estimation strategy that explicitly accounts for traders' joint inference problem increases out-of-sample return predictability at the monthly horizon by 144.3%, from $R^2 = 3.65\%$ to $R^2 = 9.35\%$, suggesting that the feature-selection problem is important to real-world traders.

JEL Classification: D83, G02, G12, G14

Keywords: Feature-Selection Risk, Sparsity, Market Dimension, Market Efficiency, Under-reaction

---

*I would like to thank Brad Barber, Roger Edelen, Xavier Gabaix, Harrison Hong, Johannes Stroebel, Jeff Wurgler, and Haoxiang Zhu for extremely helpful comments and suggestions. This paper has also benefited greatly from presentations at the AFA Annual Meeting, the Chicago Junior Finance Meeting, Illinois, Rochester, and UC Davis.

[†]University of Chicago, Booth School of Business; alexchinco@gmail.com.

Current Version: http://www.alexchinco.com/feature-selection-risk.pdf

1 Introduction

In an efficient market, if a few stocks suddenly get mispriced because they share a common feature, such as being involved in a crowded trade or getting mentioned in an M&A rumor, then fully-rational traders should rapidly exploit and eliminate this error. However, markets do not always appear efficient, and people have suggested a variety of trader shortcomings to explain why. For instance, traders might face limits to arbitrage as in Miller (1977), suffer from cognitive biases as in Daniel, Hirshleifer, and Subrahmanyam (1998), or exhibit the symptoms of non-standard preferences as in Barberis, Huang, and Santos (2001).

But, is it always right to blame traders? Perhaps a market's inefficiency has more to do with its dimensions than with its traders' limitations? After all, modern financial markets are extremely complex and densely interconnected. For any given pricing error, there are often many plausible explanations. Did Callaway Golf's stock just plunge because it happened to be involved in a crowded short-term trading strategy? Or was it because there is some truth to that new rumor about Callaway acquiring Fortune Brand? A trader should respond differently to each of these hypotheses, shorting the other stocks in the crowded strategy in the first case and buying shares of Fortune in the second. In a high-dimensional setting where assets can share many overlapping features, i.e., where Callaway can be both involved in a crowded trade and mentioned in an M&A rumor, markets do not always provide enough information to sort through the many competing hypotheses.

This paper shows that, when traders have to simultaneously decide which features are mispriced and how they should be correctly valued, the risk of selecting the wrong features can spill over and distort how they value assets. The high-dimensional nature of financial markets can act like a cognitive constraint even if traders are fully rational.

Imagine you are a trader, and each company's stock returns can have exposure to any combination of 7 features: #1) whether the company has been involved in a crowded trade (Khandani and Lo, 2007), #2) whether it has been mentioned in a news article about M&A activity (D'Aspremont and Luss, 2012), #3) whether there has been an announcement about its major supplier (Cohen and Frazzini, 2008), #4) whether its labor force has recently unionized (Klasa, Maxwell, and Ortiz-Molina, 2009), #5) whether the company belongs to the tobacco industry (Hong and Kacperczyk, 2009), #6) whether it has been referenced in a scientific journal article (Huberman and Regev, 2001), and #7) whether the company has recently been added to the S&P 500 (Barberis, Shleifer, and Wurgler, 2005). Moreover, suppose you have a hunch that there has been a shock to one of these features, but you do not initially know which one. All you know is that the market has not fully appreciated the shock, and stocks with this mystery feature will realize abnormal returns of $\alpha > 0$ in the near future. How many stock returns do you need to see to figure out which, if any, of the shocks has occurred? Three.

Suppose the first company has exposure to features $\{1, 3, 5, 7\}$—i.e., it is involved in a crowded trade, there has been an announcement about its major supplier, it belongs to the tobacco industry, and the company has been recently added to the S&P 500. Similarly, suppose that the second company has exposure to features $\{2, 3, 6, 7\}$ and the third company has exposure to features $\{4, 5, 6, 7\}$. The abnormal returns for these three stocks always reveals exactly which feature-specific shock has occurred. For instance, if only the first stock has positive abnormal returns, $ar_1 = \alpha$ while $ar_2 = ar_3 = 0$, then it must have been a crowded trade:

$$\begin{bmatrix} ar_1 \\ ar_2 \\ ar_3 \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{1}$$

Whereas, if both the first and the third stock have positive abnormal returns, $ar_1 = ar_3 = \alpha$ while $ar_2 = 0$, then it must have been a shock to the tobacco industry. There is no way to identify which feature-specific shock has occurred using fewer stocks. You need at least 8 bits of information to pick out which of the 7 feature-specific shocks has occurred and rule out the possibility of no change and $2^3 = 8$.

Now, to see how this relates to market efficiency, rewind the clock and consider the problem you face after seeing only the first two companies' abnormal returns, $ar_1 = \alpha$ and $ar_2 = 0$:

$$\begin{bmatrix} \alpha \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} ? \\ ? \\ \vdots \\ ? \end{bmatrix} \tag{2}$$

Since the first and the fifth columns are $[\,1\ \ 0\,]^\top$, you know that either a crowded trade has occurred or a tobacco-industry shock has occurred. It has to be one of these two features since these are the only two features that the first company has exposure to but the second company does not. So, what is the right way to value the third company's stock which has exposure to features $\{4, 5, 6, 7\}$, meaning that it is in the tobacco industry but not involved in a crowded trade?

There are two possibilities. In the event of a crowded trade, you should leave the third company's value unchanged; whereas, if there was a tobacco-industry shock, then you should revise your valuation. Thus, after seeing only two observations, you have to split the difference. If it was in fact the tobacco-industry shock, then you will only update halfway, and it will look like you were slow to react to public information. By contrast, if it was a crowded trade, then when you revise your valuation of the third company's stock halfway, it will look like you were over-reacting to noise. Nevertheless, this is the best you can do in real time. It is not like you are making some cognitive error or fighting against some trading friction. Instead, it is the dimensionality of your inference problem that is generating the extra risk, that is warping your perception of the third company's value, that is distorting prices.

Of course, this is just a stylized example. There are only a handful of assets, each asset's feature exposures are hand-picked, and their fundamental values do not reflect standard risk factors. To address these concerns, I apply tools from the compressed-sensing literature to generalize the result and show that if traders have seen fewer than $N^\star(Q, K)$ observations[1]

$$N^\star(Q, K) \asymp K \cdot \log(Q/K) \tag{3}$$

then, no matter what inference strategy they use, traders cannot always identify which features have realized a shock in a large market with an arbitrary number of features, Q, and an arbitrary number of shocks, K. This feature-selection bound holds even when feature exposures are randomly assigned and when companies' fundamental values reflect the usual risk factors. Moreover, in the presence of noise, some feature-selection risk will remain even after the bound has been reached. Thus, feature-selection risk is

[1] $f_N \asymp g_N$ denotes asymptotically bounded above and below, implying both $f_N = \mathrm{O}(g_N)$ and $g_N = \mathrm{O}(f_N)$.

endemic to any high-dimensional market. If assets share many overlapping features, then markets might not provide enough information to pinpoint exactly which ones matter. It is as if the high-dimensional nature of the market is introducing a cognitive constraint even though the traders themselves are fully rational.

In order to quantify the extent to which feature-selection risk limits market efficiency, I study a Kyle (1985)-type model with N assets whose values are a function of $K \ll Q$ feature-specific shocks. The model is entirely standard except for one key detail: uninformed traders, like the market maker and any would-be arbitrageurs, do not know ahead of time which K feature-specific shocks to analyze. For instance, if there is a lot of demand for Callaway Golf, the market maker now has to ask herself: Is this demand saying something about a feature-specific shock? If so, which one? Or did Callaway just happen to realize a large noise-trader demand shock? By adding this simple twist, it is possible to extend the standard information-based asset-pricing models to allow for feature-selection risk. When uninformed traders now have to infer both the identity and the size of the $K \ll Q$ feature-specific shocks, they are less responsive to aggregate demand shocks than in the original Kyle (1985) setup, making equilibrium prices less accurate.

If the market maker has to sort through a sufficiently large number of potentially relevant features, then she is likely to make a feature-selection error, regardless of the volatility of noise-trader demand. What's more, real-world traders often try to exploit this fact. For example, the co-CEO of Renaissance Technologies, Robert Mercer, has pointed out that "some signals that make no intuitive sense do indeed work. . . The signals that we have been trading without interruption for 15 years make no sense, otherwise someone else would have found them."[2] Such signals are hidden in a large feature space rather than behind a lot of noise-trader demand volatility.

Finally, I give empirical evidence that real-world traders actually care about solving this joint inference problem. To do this, I collect data on 79 different monthly factors that have been used in the asset-pricing literature. Then, for each NYSE stock I analyze 24-month rolling windows from January 1990 to December 2010, estimate the $K \ll 79$ important factor loadings using a penalized-regression procedure, and predict the stock's excess return in the subsequent month. This penalized regression

[2]Mallaby, S. (2010) *More Money Than God* (1 ed.) Penguin Books.

sets all of the "smaller" factor loadings to zero and makes it possible to estimate a sparse subset of the 79 coefficients using only 24 months of data, an approach that would clearly be unidentified using the standard regression techniques. Using this estimation strategy that explicitly accounts for traders' feature-selection problem increases the accuracy of out-of-sample return predictions at the monthly horizon by 144.3%, from $R^2 = 3.65\%$ to $R^2 = 9.35\%$! Thus, solving this joint inference problem is very important for real-world traders.

1.1 Related Literature

This paper borrows from and brings together several strands of literature. First, the current paper is closely related to the literature on bounded rationality; yet, there is a fundamental difference in approaches. Existing theories use cognitive constraints to induce boundedly rational decision making. For example, papers like Sims (2006) and Hong, Stein, and Yu (2007) suggest that cognitive costs force traders to use overly simplified mental models, and Gabaix (2013) derives the sort of mental models that traders would choose when facing ℓ_1 thinking costs. By contrast, I use bandwidth constraints on a market's *signals* rather than on a trader's *processing power* to generate similar behavior. Both channels are at work in asset markets. This paper is the first to articulate the bandwidth constraint on a finite set of market signals. To do this, I use the results from the compressed sensing literature, which originated with Candes and Tao (2005) and Donoho (2006).

Second, the model relies on the fact that asset values are governed at least in part by a constantly changing cast of feature-specific shocks. Chinco, Clark-Joseph, and Ye (2019) provides evidence both that assets realize many different kinds of characteristic-specific shocks and also that it is hard for traders to identify which ones are relevant in real time. This assumption is consistent with, but separate from, existing asset-pricing models. On the theoretical side, it is possible to fit this high-dimensional problem into many popular asset-pricing models since they contain substantial amounts of theoretical "dark matter" in the language of Chen, Dou, and Kogan (2014).

On the empirical side, the high-dimensional and ever-changing nature of trader's problem has been documented in a series of papers on data-snooping. For a representative sample, see Lo and MacKinlay (1990), Sullivan, Timmermann, and White (1999), and Kogan and Tian (2014). In particular, Kogan and Tian (2014) notes that parameter estimates for factor loadings are "highly sensitive to the sample period choice and the details of the factor construction. In particular, there is virtually no correlation between the relative model performance in the first and the second halves of the 1971-2011 sample period. Using a two-way sort on firm stock market capitalization (size) and characteristics to construct model return factors, an often used empirical procedure, similarly scrambles the relative model rankings."

Campbell, Lettau, Malkiel, and Xu (2001) also give evidence that the usual factor models only account for a fraction of firm-specific return volatility. For example, if you selected an NYSE/AMEX/NASDAQ stock at random in 1999, market and industry factors only accounted for 30% of the variation in its daily returns. Recent work by Ang, Hodrick, Xing, and Zhang (2006), Chen and Petkova (2012), and Herskovic, Kelly, Lustig, and Van Nieuwerburgh (2014) gives strong evidence that there is a lot of cross-sectional structure in the remaining 70% of so-called idiosyncratic volatility—i.e., patterns in past idiosyncratic volatility are strong predictors of future returns. Thus, some portion of the 70% remainder appears to be neither permanent factor exposure nor fully idiosyncratic events.

Finally, this paper also gives a mathematical foundation for Hayek (1945)'s notion of local knowledge. A trader who benefits from specialized experience with particular assets is the canonical example of local knowledge. One way to interpret the results is as something of an anti-Harsanyi doctrine and a microfoundation for the behavioral finance literatures on disagreement (Hong and Stein, 2007) and noise trading (Black, 1986). This paper gives a situation where 2 rational Bayesian market makers can look at the exact same aggregate demand schedules for $N < N^{\star}(Q, K)$ assets and not have the same posterior beliefs due to the dimensionality of the problem.

2 Baseline Equilibrium Model

I begin by characterizing a baseline equilibrium where traders do not face any feature-selection risk. Specifically, I assume that they have access to an oracle alerting them to the K features that have realized a shock but not the size or sign of this shock. I will return to this baseline model later as a point to comparison to answer the question: how does feature-selection risk alter the usual predictions?

2.1 Market Structure

I study a market containing N risky assets with fundamental values, v_n, governed by exposure to Q features:

$$v_n = \sum_{q=1}^{Q} \alpha_q \cdot x_{n,q} \tag{4}$$

where $x_{n,q} \overset{\text{iid}}{\sim} \text{Normal}(0,1)$ denotes asset n's exposure to the qth feature and α_q denotes the size of the shock to feature q. So, for example, if there is a shock of size $\alpha_{\text{Tobacco}} = \$1$ to stocks in the tobacco industry, then the share price of a company in that industry, $x_{n,\text{Tobacco}} = 1$, will rise by \$1.

I consider the setting where everyone knows each asset's feature exposures $\boldsymbol{x}_n$. i.e., all agents have a detailed list of whether or not each asset has been involved in a crowded trade, mentioned in an article on M&A activity, suffered a setback to one of its suppliers, etc... If there is any uncertainty in later sections, it will be about which elements in $\boldsymbol{\alpha}$ are non-zero. For instance, traders might be uncertain about whether or not the tobacco industry has realized a shock, but they will never be uncertain about whether a particular company is in the tobacco industry.

Because each asset has different feature exposures, each asset will manifest a feature-specific shock in a slightly different way. For example, we know that some stocks are more likely to be included in statistical arbitrage strategies than others, news of M&A activity has an opposite effect on the acquirer and the target, and some companies are more strongly impacted by news about a particular supplier than others. Suppose that asset 1 has exposures to the stat-arb-strategy, M&A activity, and supplier stock features given by $\boldsymbol{x}_1 = [\, 1.50 \;\, 0.50 \;\, -0.10 \,]^\top$ while asset 2 has feature exposures $\boldsymbol{x}_2 = [\, -0.50 \;\, -0.75 \;\, 1.00 \,]^\top$. Each stock's value will then be:

$$v_1 = \alpha_{\text{StatArb}} \times (+1.50) + \alpha_{\text{M\&A}} \times (+0.50) + \alpha_{\text{EconLink}} \times (-0.10) + \cdots \tag{5a}$$

$$v_2 = \alpha_{\text{StatArb}} \times (-0.50) + \alpha_{\text{M\&A}} \times (-0.75) + \alpha_{\text{EconLink}} \times (+1.00) + \cdots \tag{5b}$$

Thus, a positive M&A activity shock of $\alpha_{\text{M\&A}} = 1$ will lead to a \$0.50 rise in the fundamental value of asset 1. By contrast, the same shock will lead to a \$0.75 decline in the fundamental value of asset 2. Same shock. Different feature exposures. Opposite affects on value.

To capture the idea that only a few of the many possible features that might impact a stock's value each period actually matter, I study a world where only K of the elements in $\boldsymbol{\alpha}$ are non-zero

$$K = \|\boldsymbol{\alpha}\|_0 = \sum_{q=1}^{Q} 1_{\{\alpha_q \neq 0\}} \tag{6}$$

with $Q \gg N \geq K$ and $\mathcal{K}$ denoting the set of shocked features. I also assume the vector of feature-specific shocks, $\boldsymbol{\alpha}$, satisfies the following conditions:

1. $\mathcal{K} \subset \{1, 2, \ldots, Q\}$ is selected uniformly at random.
2. The signs of $\boldsymbol{\alpha}_{[\mathcal{K}]}$ are independent and equally likely to be -1 or $+1$.
3. The magnitudes of $\boldsymbol{\alpha}_{[\mathcal{K}]}$ are independent and bounded by $\alpha_{\max} \geq |\alpha_q| > \sigma_z$.

To be sure, shocks are never really exactly sparse; they are only approximately sparse meaning that they may be well approximated by sparse expansions. All of the results in this paper go through if you assume that K features realize shocks that are much bigger than the rest.

This market structure means that it is possible for a trader to see several assets behaving wildly without being able to put his finger on which K feature-specific shocks are the culprit. For instance, the chairman of Caxton Management, Bruce Kovner, notes that there are often many plausible reasons why prices might move in either direction at any point in time. "During the past six months, I had good arguments for the Canadian dollar going down, and good arguments for the Canadian dollar going up. It was unclear to me which interpretation was correct."[3] This was not a situation where Kovner had to learn more about a well-defined trading opportunity; rather, the challenge was to pick which explanation to trade on in the first place. Kovner faced feature-selection risk.

Of course, sometimes traders are not in the business of spotting feature-specific shocks. For example, a January 2008 Chicago Tribune article about Priceline.com reported that "a third-quarter earnings surprise sent [the company's] shares skyward in November, following an earlier announcement that the online travel agency planned to make permanent a no-booking-fees promotion on its airline ticket purchases."[4] No

[3]Schwager, J. (1989) *Market Wizards: Interviews with Top Traders.* (1 ed.) New York Institute of Finance.

[4]DiColo, J. (2008, Jan. 20) Priceline's Power Looks Promising in Europe, Asia. *Chicago Tribune*.

one was confused about why Priceline's price rose. The only problem facing traders was deciding how much to adjust the price. Existing information-based asset-pricing models are well suited to this setting.

2.2 Objective Functions

There are two kinds of optimizing agents, asset-specific informed traders and a market-wide market maker, along with a collection of asset-specific noise traders.

Asset-specific informed traders know the fundamental value of a single asset, v_n, and solve the standard static Kyle (1985)-type optimization problem with risk neutral preferences,

$$\max_{y} \mathbb{E}\left[\, (v_n - p_n) \cdot y \,|\, v_n \,\right] \tag{7}$$

where y denotes the size of asset n's informed trader's market order in units of shares and p_n denotes the price that they pay in units of dollars per share. Crucially, for these traders, the fundamental value of each asset is just a random variable with no further structure. They do not observe which K feature-specific shocks govern its value.

There are a couple of ways to justify this assumption. First, you might think about the asset-specific informed traders as value investors. For instance, Li Lu, founder of Himalaya Capital and well known value investor, suggests that in order to gain market insight you should "Pick one business. Any business. And truly understand it. I tell my interns to work through this exercise—imagine a distant relative passes away and you find out that you have inherited 100% of a business they owned. What are you going to do about it?"[5] It is like they have a gut instinct. Alternatively, you can think about the informed traders as getting a signal about the level of noise-trader demand in a given asset. They can then invert this information about noise-trader demand to learn something about fundamentals, v_n, without learning its underlying structure.

The market-wide market maker observes aggregate demand, d_n, for each asset

$$d_n = y_n + z_n \tag{8}$$

which is composed of demand from asset-specific informed traders, y_n, and from noise traders, $z_n \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_z^2)$. He then tries to set the price of each asset as close as

[5]Lu, L. (2010) *Lecture at Columbia Business School.*

possible to its fundamental value given the cross-section of aggregate demand:

$$\min_{\boldsymbol{p}} \mathbb{E}\left(\tfrac{1}{N} \cdot \textstyle\sum_{n=1}^{N} (p_n - v_n)^2 \,\middle|\, \boldsymbol{d} \right) \tag{9}$$

Put differently, competitive pressures force the market maker to try and minimize the mean squared error between the price and each asset's value.

Notice that this formulation of the market maker's problem is slightly different from the one in the original Kyle (1985) model. Here, the market maker explicitly minimizes his prediction error; whereas, in the original setup, the market maker just sets the price equal to his conditional expectation, which happens to minimize his prediction error since there are as many assets as shocks. In the current paper, it is important that the market maker explicitly minimizes his prediction error because the conditional expectation will no longer be well defined when there are more possible feature-specific shocks than assets.

Because there are many more features than assets, $Q \gg N \geq K$, the market maker must use a feature-selection rule $\phi(\boldsymbol{d}, \boldsymbol{X})$ that accepts an $(N \times 1)$-dimensional vector of aggregate demand as well as an $(N \times Q)$-dimensional matrix of features and then returns a $(Q \times 1)$-dimensional vector of estimated feature-specific shocks:

$$\phi : \boldsymbol{R}^N \times \boldsymbol{R}^{N \times Q} \mapsto \boldsymbol{R}^Q \tag{10}$$

$\hat{\boldsymbol{\alpha}} \overset{\text{def}}{=} \phi(\boldsymbol{d}, \boldsymbol{X})$ denotes the estimated shocks. Later, I will give bounds on how well the best possible feature-selection rule can perform in a market with Q features, K shocks, and N assets. The nature of the equilibrium asset prices will depend on how much information about the sparse feature-specific shocks, $\boldsymbol{\alpha}$, the market maker can tease out of the cross-section of aggregate demand, $\boldsymbol{d}$. It is clear that real-world traders worry about how much their market maker can learn from the combination of their orders. Quant hedge funds place orders for different legs of the same trade with different brokers to make it difficult for brokers to do this sort of reverse engineering.

2.3 Oracle Equilibrium

Let's now explore the equilibrium when the market maker has an oracle telling him exactly which K features have realized a shock. It turns out that the coefficients in

Proposition 2.3 below are identical to the standard Kyle (1985) model coefficients. This fact highlights how existing information-based asset-pricing models implicitly assume that all traders know exactly which features to study.

Figure 1 summarizes the timing of the model. First, nature assigns feature exposures to the N assets and picks a subset of K features to realize shocks. After the exposures and shocks have been drawn but before any trading takes place, the N asset-specific informed traders learn the fundamental value of their own asset, v_n, and the single market maker common to all N assets observes which K features have realized a shock (but not the size or sign of these shocks). Finally, trading takes place. Each of the N informed traders and noise traders places a market order. Then, the market maker observes each asset's aggregate order flow, updates his conditional expectation about their values, and sets prices accordingly.

An equilibrium, $\mathcal{E} \overset{\text{def}}{=} \{\theta, \lambda\}$, is a linear demand rule for each of the N asset-specific informed traders, $y_n = \theta \cdot v_n$, and a linear pricing rule for the single market maker common to all N assets, $p_n = \lambda \cdot d_n$, such that (a) the demand rule θ solves Equation (7) given the correct assumption about λ and (b) the pricing rule λ solves Equation (9) given the correct assumption about θ.

Proposition 2.3 (Oracle Equilibrium)**.** *If the market maker knows $\mathcal{K}$, then there exists an equilibrium defined by coefficients:*

$$\lambda = 1/(2 \cdot \theta) \tag{11a}$$

$$\theta = \sqrt{K/N} \cdot (\sigma_z/\sigma_v) \tag{11b}$$

Because there are more assets than feature-specific shocks, the market maker can just run the standard OLS regression

$$d_n/\theta = \boldsymbol{x}_n \, \hat{\boldsymbol{\alpha}}_{\text{OLS}} + \epsilon_n \tag{12}$$

to estimate $\hat{\boldsymbol{\alpha}}_{\text{OLS}}$. Knowing these coefficients then gives him an unbiased signal, $\boldsymbol{X} \, \hat{\boldsymbol{\alpha}}_{\text{OLS}}$, about each asset's fundamental value. This signal has variance

$$\mathbb{E}\Big[\, \|\boldsymbol{v} - \boldsymbol{X} \, \hat{\boldsymbol{\alpha}}_{\text{OLS}}\|_2^2 \, / \, N \, \Big] = K \cdot (\sigma_z^2/\theta^2) \, / \, N \tag{13}$$

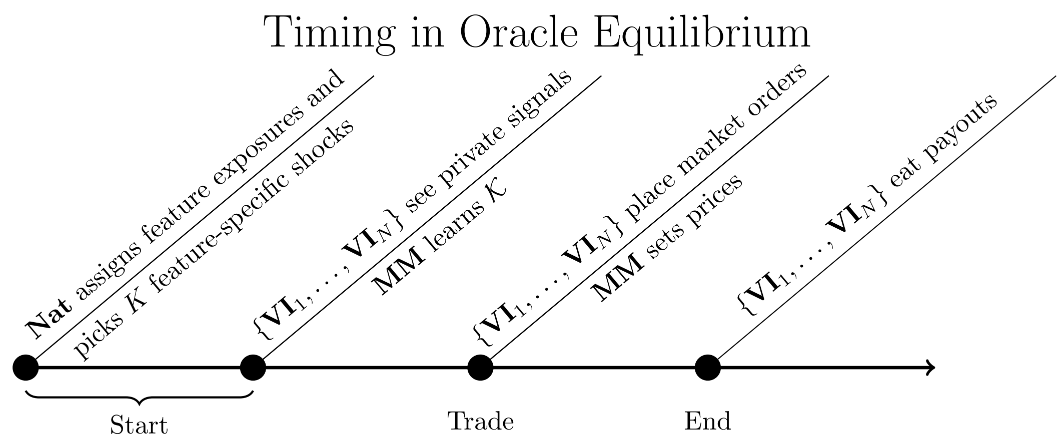


Figure 1: What each agent knows and when they know it in the model where the common market maker knows which K features have realized a shock. "Nat" is nature. "VI" is the value investor. "MM" is the market maker.

where $\boldsymbol{v}$ is an $(N \times 1)$-dimensional vector of asset values. Using his priors on the distribution of each asset's value, $v_n \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_v^2)$, he can then use DeGroot (1969) updating to form posterior beliefs. The market maker's signal error is increasing in the variance of noise-trader demand, so he has a harder time figuring out if a positive demand shock is due to noise traders or just a really strong fundamental value realization. Thus, more noise-trader demand volatility means informed traders have an easier time masking their trades allowing them to trade more intensely.

3 Feature-Selection Bound

We just saw what the equilibrium looks like when traders know exactly which features to analyze. Let's now look at how hard it is to recover this information without an oracle. Specifically, I show that, if the market maker has not seen at least $N^\star(Q, K)$ observations, then he will always suffer from feature-selection risk and will always make some errors in picking which features to analyze, no matter what inference strategy he uses and even if he is fully rational.

3.1 Theoretical Minimum

Suppose that the market maker was the most sophisticated trader ever and could choose the best inference strategy possible, ϕ_{Best}. How many observations does he need to see to be sure that he has identified which feature-specific shocks have taken place? He does not need to see Q observations since the vector $\boldsymbol{\alpha}$ is K-sparse. But what is this bare minimum number?

To answer this question, I consider limiting results for sequences of markets $\{(Q_N, K_N)\}_{N \geq 0}$ where the number of features, $Q = Q_N$, and the sparsity level, $K = K_N$, are allowed to grow with the number of observations, N:

$$\lim_{N \to \infty} Q_N, K_N = \infty \qquad\qquad \lim_{N \to \infty} K_N / Q_N = 0 \tag{14}$$

For example, take $K = \sqrt{Q}$. This asymptotic formulation captures the spirit of traders' joint inference problem. For instance, Daniel (2009) notes that during the Quant Meltdown of August 2007 "markets appeared calm to non-quantitative investors…you could not tell that anything was happening without quant goggles" even though large funds like Highbridge Capital Management were suffering losses on the order of 16%.[6] All stocks with exposure to the held-in-a-stat-arb-strategy feature realized a massive shock, but this feature was just one of many plausible feature-specific shocks that might have occurred ex ante. Unless you knew where to look (had "quant goggles"), the event just looked like noise.

I define the market maker's feature-selection error as the quantity

$$\mathrm{FSE}[\phi] = \mathbb{E}\left[\, \|\mathrm{S}[\hat{\boldsymbol{\alpha}}] - \mathrm{S}[\boldsymbol{\alpha}]\|_{\infty} \,\right] \tag{15}$$

where the operator $\mathrm{S}[\cdot]$ identifies the support of a vector:

$$\mathrm{S}[\hat{\alpha}_q] = \begin{cases} 1 & \text{if } \hat{\alpha}_q \neq 0 \\ 0 & \text{if } \hat{\alpha}_q = 0 \end{cases} \tag{16}$$

The ℓ_∞-norm gives a 1 if there is any difference in the support of the vectors and a 0 otherwise. In words, $\mathrm{FSE}[\phi]$ is the probability that the market maker's selection rule, ϕ, chooses the wrong subset of features when averaging over not only the measurement noise but also the choice of the Gaussian exposure matrix, $\boldsymbol{X}$. Let $\boldsymbol{\Phi}$ denote the set of all possible inference strategies the market maker might use. If there exists some

[6] Zuckerman, G., J. Hagerty, and D. Gauthier-Villars (2007) Mortgage Crisis Spreads. *Wall Street Journal*.

inference strategy $\phi \in \mathbf{\Phi}$ with $\mathrm{FSE}[\phi] = 0$, the market maker can use this approach to always select which feature-specific shocks have taken place with probability 1—i.e., there exists (at least in principle) an inference strategy that would be just as good as having an oracle. It may not be computationally feasible, but it would exist.

The feature-selection bound in Proposition 3.1 below then says that no such strategy exists when the market maker has seen fewer than $N^{\star}(Q, K)$ observations. When $N < N^{\star}(Q, K)$, at least a few feature-selection errors are unavoidable regardless of what approach $\phi \in \mathbf{\Phi}$ the market maker takes.

Proposition 3.1 (Feature-Selection Bound)**.** *If there exists a constant* $C > 0$ *such that*

$$N < C \cdot K_N \cdot \log(Q_N / K_N) \tag{17}$$

as $N \to \infty$*, then there exists some constant* $c > 0$ *such that*

$$\min_{\phi \in \mathbf{\Phi}} \mathrm{FSE}[\phi] > c \tag{18}$$

The threshold value $N^{\star}(Q, K) \asymp K \cdot \log(Q/K)$ *is the feature-selection bound.*

Importantly, Proposition 3.1 does not make any assumptions about the market maker's cognitive abilities. It says that when $N < N^{\star}(Q, K)$ the market maker has to be misinterpreting aggregate-demand signals at least some of the time due to the nature of his sparse, high-dimensional, inference problem. Put another way, this minimum number of observations is a consequence of a theoretical bound on how informative market signals can be rather than a consequence of thinking costs or trading frictions. In some sense, it has nothing to do with the market maker. He could be Einstein, Friedman, and Kasparov all rolled into one and it would not matter. There is simply a lower bound on the amount of data needed to say anything useful about which market events have taken place using the cross-section of aggregate demand. This is a very different way of thinking about why rational traders sometimes misinterpret market signals. This result is first derived in Wainwright (2009a).

3.2 Discussion

There are a couple of points about the interpretation of Proposition 3.1 worth discussing in more detail. First, while the asymptotics are helpful for analytical reasons, they are not critical to the underlying result. There is a qualitative change in the nature of any inference problem when you move from choosing which feature-specific shocks have occurred to deciding how large they must have been. To see why, let's return to the example in Section 1 where only 1 of the 7 features might have realized a shock, and consider the more general case where any of the 7 features could have. This gives

$$\begin{aligned} 128 &= \binom{7}{0} + \binom{7}{1} + \binom{7}{2} + \binom{7}{3} + \binom{7}{4} + \binom{7}{5} + \binom{7}{6} + \binom{7}{7} \\ &= 1 + 7 + 21 + 35 + 35 + 21 + 7 + 1 \\ &= 2^7 \end{aligned} \tag{19}$$

different feature combinations. Thus, $N^{\star}(7, 7) = 7$ provides a trader just enough differences to identify which combination of features has realized a shock. More generally, for any number of features, Q, a trader needs $2^Q = \sum_{k=0}^{Q} \binom{Q}{k}$ observations to detect shocks if he has no information about K. This gives an information theoretic interpretation to the meaning of "just identified" that has nothing to do with linear algebra or matrix invertibility.

Second, these asymptotics do not pose a practical problem when applying the bound. To begin with, real-world markets are finite but very large, so the asymptotic approximation is a good one. While it is not possible to give a precise formulation of the feature-selection bound in the finite sample case, practical compressed sensing techniques can make error-rate guarantees in finite samples. What's more, analysts regularly make this sort of asymptotic-to-finite leap in mainstream econometric applications. For example, the practical implementation of GMM involves a 2-step procedure as outlined in Newey and McFadden (1994). The first step estimates the coefficient vector using the identity weighting matrix on the basis that any positive-semidefinite weighting matrix will give the same point estimate in the large T limit. The second step uses the realized point estimates to compute the standard errors.

Finally, the result in Proposition 3.1 is likely too optimistic about the ability of the most sophisticated market maker since it makes no assumptions about the inference strategy being convex. How much harder could the non-convex approach be? A lot. Consider the example from the Introduction. Suppose that $Q = 400$, and I told you that exactly $K = 5$ of the features were mispriced. You can certainly try to solve the general problem by tackling each of the $\binom{400}{5} \approx 8.3 \times 10^{10}$ sub-problems with a regression procedure. This is a huge number of cases to check on par with the number of bits in the human genome. As Rockafellar (1993) writes, "the great watershed in optimization isn't between linearity and nonlinearity, but convexity and non-convexity."

4 Feature-Selection Risk

If the market maker has not seen at least $N^{\star}(Q, K)$ observations, then he will always suffer from feature-selection risk no matter what inference strategy he uses. Let's now introduce this feature-selection risk into the baseline asset-pricing model to see how it warps the market maker's perception of asset values. The basic equilibrium concept will remain completely standard. The key question to ask is: how much information about which K feature-specific shocks have occurred can the market maker infer from the cross-section of aggregate demand of N stocks?

4.1 Inference Strategy

To solve for equilibrium asset prices, I need to compute the market maker's posterior beliefs after observing the cross-section of aggregate demand, $\boldsymbol{d}$. This means I need to make a choice about which inference strategy the market maker uses.

I study a market maker who uses the least absolute shrinkage and selection operator (LASSO) outlined in Tibshirani (1996)

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\tilde{\boldsymbol{\alpha}}} \left\{ \| \boldsymbol{X} \tilde{\boldsymbol{\alpha}} - \boldsymbol{d}/\theta \|_2^2 + \gamma \cdot \|\tilde{\boldsymbol{\alpha}}\|_{\ell_1} \right\} \tag{20}$$

for $\gamma > 0$. The ℓ_1 norm means that the LASSO sets all coefficient estimates with $|\alpha_q| < \gamma$ equal to zero. It generates a preference for sparsity. For example, if there were no $\gamma \cdot \|\tilde{\boldsymbol{\alpha}}\|_{\ell_1}$ term, then the inference strategy would be equivalent to ordinary least squares which is not well-posed for $Q \gg N$. The tuning parameter, γ, controls how likely the estimation procedure is to get false positives. To screen out spurious variables, you want γ to be large; however, increasing γ also means that you are more likely to ignore meaningful variables that happen to look small in the data by chance. Decreasing γ to reduce this problem floods the results with spurious coefficients.

Note that in the current paper, the use of the ℓ_1-norm is not a consequence of bounded rationality as in Gabaix (2013). Rather, it is simply a way for the market maker to draw an inference about the value of each asset given the cross-section of aggregate demand. Since the market maker does not have access to an oracle, there are now more features than stocks, $Q \gg N$. As a result, his inference procedure needs to have a preference for sparsity. Any penalty with a norm $p \in [0, 1]$ will do. For example, think about the ℓ_0 problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\tilde{\boldsymbol{\alpha}}} \left\{ \| \boldsymbol{X} \tilde{\boldsymbol{\alpha}} - \boldsymbol{d}/\theta \|_2^2 + \gamma \cdot \|\tilde{\boldsymbol{\alpha}}\|_{\ell_0} \right\} \tag{21}$$

However, a penalty with a norm $p \in [0, 1)$ generates a non-convex inference problem which is computationally intractable. Natarajan (1995) explicitly shows that ℓ_0 constrained programming is NP-hard. Thus, the ℓ_1 norm, which sits right on the boundary of the two regions, is the natural choice for the penalty. What's more, when feature exposures are drawn independently from identical Gaussian distributions as they are in the current paper, the LASSO comes within a logarithmic factor of optimality as shown in Wainwright (2009b).

4.2 Equilibrium Using the LASSO

We can now solve for the equilibrium coefficients in the more general setting where the market maker does not have access to an oracle and must solve a sparse, high-dimensional, inference problem on his own. I show that informed traders in this new model earn higher profits since they can hide behind both noise-trader demand volatility and feature-selection risk.

Candes and Plan (2009) prove that if the market maker sees the aggregate demand for at least $N^\star(Q, K)$ assets, then the LASSO gives a signal about each asset's value, $\boldsymbol{v}$, with a signal error that satisfies the inequality below

$$\| \boldsymbol{X} \hat{\boldsymbol{\alpha}}_{\text{LASSO}} - \boldsymbol{v} \|_2^2 / N \leq \tilde{C}^2 \cdot \log(Q) \times K \cdot (\sigma_z^2/\theta^2) / N \tag{22}$$

with probability approaching unity as $N \to \infty$ for $\tilde{C} = 2 \cdot \sqrt{2} \cdot (1 + \sqrt{2})$. Where does this $\tilde{C}^2 \cdot \log(Q)$ factor come from? Because the market maker has to simultaneously decide both *which* asset features have realized a shock and also *how large* they were, he will sometimes make errors in identifying which features have realized a shock. When he does so, there will be additional noise in his posterior beliefs about each asset's fundamental value. It is these feature-selection errors that increase the variance of his posterior beliefs by a factor $\tilde{C}^2 \cdot \log(Q)$ relative to when he had an oracle.

The equilibrium concept will be the same as before. An equilibrium, $\mathcal{E}_\phi = \{\theta, \lambda\}$, is a linear demand rule for each of the N asset-specific informed traders, $y_n = \theta \cdot v_n$, and a linear pricing rule for the single market maker common to all N assets, $p_n = \lambda \cdot d_n$, such that (a) the demand rule θ solves Equation (7) given the correct assumption about λ and (b) the pricing rule λ solves Equation (9) given the correct assumption about θ and assuming the market maker uses to the LASSO to solve his sparse, high-dimensional, inference problem.

Proposition 4.2 (Equilibrium Using the LASSO)**.** *If the market maker uses the LASSO with $\gamma = 2 \cdot (\sigma_z/\theta) \cdot \sqrt{2 \cdot \log(Q)}$ to identify and interpret feature-specific shocks and $N > N^\star(Q, K)$, then there exists an equilibrium defined by coefficients*

$$\lambda = 1/(2 \cdot \theta) \tag{23a}$$

$$\theta = C \cdot \sqrt{\log(Q)} \times \sqrt{K/N} \cdot (\sigma_z/\sigma_v) \tag{23b}$$

for some positive numerical constant $0 < C < \tilde{C}$.

The key takeaway from Proposition 4.2 is that increasing Q, the number of payout-relevant features that a market maker has to sort through, makes the price less responsive to demand shocks. This happens via two different channels. First, increasing Q raises the feature-selection bound, $N^\star(Q, K)$, so that the market maker has to see more assets before he can correctly identify which features have realized a shock. When there are fewer than $N^\star(Q, K)$ assets for the market maker to inspect, the LASSO does not reveal anything about which feature-specific shocks have occurred. Thus, in this regime, the common market maker effectively operates in N distinct asset markets. Each asset's demand gives him information about that particular asset's fundamental value, but he cannot extrapolate this information from one asset to the next. Second, increasing Q makes the market maker less certain about his inferences. It imposes a penalty on the precision of the market maker's posterior beliefs of $C^2 \cdot \log(Q)$ per unit of fundamental volatility. It takes time to decode market signals.

Proposition 4.2 includes a numerical constant C. The exact value of this numerical constant will depend on the distribution of the sizes of the K feature-specific shocks.

The exact value of the constant can be found numerically by bootstrap procedures—i.e., by repeatedly estimating the LASSO on sample datasets. For example, when the magnitude of the K feature-specific shocks is drawn $\alpha_q \overset{\text{iid}}{\sim} \pm\text{Unif}[1,2] \cdot (\sigma_z/\theta)$, simulations reveal that $C \approx 2 \cdot (1 + \sqrt{2}) \approx 4.82$. I make no effort to characterize this value further because it depends on the gritty details of the asset value distribution. Changing C slightly does not alter the qualitative intuition behind the impact of feature-selection risk.

5 Empirical Evidence

Feature-selection risk should only matter when traders face a joint inference problem—i.e., when traders have to simultaneously decide both which features are mispriced and how they should be correctly valued. Is there any evidence that traders actually care about this problem in the real world? Yes. Following the approach introduced in Chinco, Clark-Joseph, and Ye (2019), I show that using an estimation strategy which explicitly accounts for traders' joint inference problem increases the accuracy of out-of-sample return predictions at the monthly horizon by 144.3%, from $R^2 = 3.65\%$ to $R^2 = 9.35\%$! Thus, solving this joint inference problem is important to real-world traders.

5.1 Econometric Approach

Let's first look at what it means to say that an estimation strategy accounts for traders' joint inference problem. I begin by estimating a benchmark AR(1) specification using rolling 24-month sample periods

$$rx_{n,t} = \hat{\phi}_0 + \hat{\phi}_1 \cdot rx_{n,t-1} + \epsilon_{n,t} \tag{24}$$

like in Jegadeesh (1990). This amounts to minimizing the squared prediction error

$$\hat{\boldsymbol{\phi}} = \min_{\boldsymbol{\phi} \in \mathbf{R}^2} \left\{ \frac{1}{24} \cdot \sum_{t=1}^{24} \left(rx_{n,t} - \{\phi_0 + \phi_1 \cdot rx_{n,t-1}\} \right)^2 \right\} \tag{25}$$

which is easy to do since there are many more observations, 24, than there are coefficients, 2.

After fitting the regression coefficients, I then predict the subsequent month's returns:

$$\mathbb{E}_t[rx_{n,t+1}] = f_{n,t}^{\text{AR}(1)} = \hat{\phi}_0 + \hat{\phi}_1 \cdot rx_{n,t} \tag{26}$$

If $f_{n,t}^{\text{AR}(1)}$ is a good predictor of the realized excess return, $rx_{n,t+1}$, then traders only have to think about stock-specific considerations when predicting future returns. I measure the predictive power of $f_{n,t}^{\text{AR}(1)}$ by the R^2 of an out-of-sample regression

$$rx_{n,t+1} = \tilde{a}_n + \tilde{b}_n \cdot \left(\frac{f_{n,t}^{\text{AR}(1)} - \mu_n^{\text{AR}(1)}}{\sigma_n^{\text{AR}(1)}} \right) + e_{n,t+1} \tag{27}$$

where $\mu_n^{\text{AR}(1)}$ and $\sigma_n^{\text{AR}(1)}$ are the mean and standard deviation of the predictor $f_{n,t}^{\text{AR}(1)}$.

Next, I estimate the relationship between each stock's monthly excess returns and a large collection of Q predictors where $Q \gg 24$. Since there are more plausible predictors, Q, than months in the sample period, 24, the estimation strategy in Equation (25) is no longer valid. Instead, it is necessary to use a penalized regression. I use the least absolute shrinkage and selection operator (LASSO) as introduced in Tibshirani (1996) and used in Section 4 above. This means choosing a $([Q+1] \times 1)$-dimensional vector of coefficients using the optimization problem below:

$$\hat{\boldsymbol{\varphi}} = \min_{\boldsymbol{\varphi} \in \boldsymbol{R}^{Q+1}} \left\{ \frac{1}{2 \cdot 24} \cdot \sum_{t=1}^{24} \left(rx_{n,t} - \left\{ \varphi_0 + \sum_{q=1}^{Q} \varphi_q \cdot f_{q,t-1} \right\} \right)^2 + \lambda \cdot \sum_{q=1}^{Q} |\varphi_q| \right\} \tag{28}$$

The LASSO penalty function, $\lambda \cdot \sum_{q=1}^{Q} |\varphi_q|$, sets all OLS coefficient estimates that are smaller than $|\varphi_q| < \lambda$ to zero as discussed in Section 4 above. Thus, the LASSO both selects and estimates the relevant coefficient loadings. This is the sense in which the LASSO explicitly incorporates traders' joint inference problem.

After fitting the regression coefficients, I again predict the subsequent month's returns:

$$f_{n,t}^{\text{LASSO}} \overset{\text{def}}{=} \mathbb{E}_t[rx_{n,t+1}] = \hat{\varphi}_0 + \sum_{q=1}^{Q} \hat{\varphi}_q \cdot f_{q,t} \tag{29}$$

If the predicted $f_{n,t}^{\text{LASSO}}$ is a good predictor of the realized excess return, $rx_{n,t+1}$, then traders' joint inference problem is crucial to predicting stock returns. I measure the predictive power of $f_{n,t}^{\text{LASSO}}$ by the R^2 of an out-of-sample regression

$$rx_{n,t+1} = \tilde{a}_n + \tilde{c}_n \cdot \left(\frac{f_{n,t}^{\text{LASSO}} - \mu_n^{\text{LASSO}}}{\sigma_n^{\text{LASSO}}} \right) + e_{n,t+1} \tag{30}$$

where μ_n^{LASSO} and σ_n^{LASSO} are the mean and standard deviation of the predictor $f_{n,t}^{\text{LASSO}}$.

To make sure that both predictors are not capturing the exact same information, I also run a regression with both predictors on the right-hand side:

$$rx_{n,t+1} = \tilde{a}_n + \tilde{b}_n \cdot \left(\frac{f_{n,t}^{\text{AR}(1)} - \mu_n^{\text{AR}(1)}}{\sigma_n^{\text{AR}(1)}} \right) + \tilde{c}_n \cdot \left(\frac{f_{n,t}^{\text{LASSO}} - \mu_n^{\text{LASSO}}}{\sigma_n^{\text{LASSO}}} \right) + e_{n,t+1} \tag{31}$$

The logic here is simple. If the stock-specific information captured by the AR(1) model and feature-selection information captured by the LASSO model are really different kinds of information, then the R^2 of this combined regression will be roughly equal to the sum of the R^2s from Equations (27) and (30).

5.2 Data Sources

I collect 79 different predictive variables at the monthly horizon from January 1990 to December 2010 from a variety of data sources. Monthly returns for NYSE stocks come from the Wharton Research Data Service (WRDS).

The bulk of the predictors come from Ken French's website. See `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` for more detailed variable definitions. I include factors representing the excess returns to the market portfolio as well as portfolios of small, large, growth, and value stocks as in Fama and French (1993). I also consider factors representing the excess returns to medium-term momentum (Jegadeesh and Titman, 1993) as well as to short- and long-term reversals (Jegadeesh, 1990). In addition, there are factors representing the excess returns to portfolios of high and low operating profit firms and high and low real-investment firms. Table 1a houses summary statistics for all of these predictors.

The same data library also contains data on the monthly excess returns to country- and industry-specific portfolios. I include factors for the following countries: Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Greece, Hong Kong, Ireland, Italy, Japan, Netherlands, New Zealand, Norway, Portugal, Singapore, Spain, Switzerland, Sweden, and United Kingdom. Table 1c houses the summary statistics for these factors. The absence of many of the country-specific factors prior to January 1990 dictates the starting point for my sample period. Similarly, I include factors for 30 different SIC Code industries. See Table 1d for the relevant descriptive statistics.

I incorporate data on a variety of market-sentiment indicators used in Baker and Wurgler (2006). These data are all available on Jeff Wurgler's website: `http://people.stern.nyu.edu/jwurgler/`. The dividend premium originally comes from Baker and Wurgler (2004), the number and first-day return on IPOs is defined in Ibbotson, Sindelar, and Ritter (1994), the average monthly turnover of NYSE stocks comes from the NYSE Factbook, the closed-end-fund discount is detailed in Neal and Wheatley (1998), and the equity share in new issues is originally outlined in Baker and Wurgler (2000). The sentiment index is a factor representing the first principal component of these six sentiment proxies over 1962-2005 time period, where each of the proxies has first been orthogonalized with respect to a set of macroeconomic conditions. Table 1b displays the relevant summary statistics.

I also add a variety of other macroeconomic predictors: a recession indicator as defined by the National Bureau of Economic Research (NBER), factors representing the U.S. employment growth rate and the U.S. inflation rate from the Bureau of Economic Analysis (BEA), and a factor denoting the level of the VIX from the Chicago Board of Options Exchange (CBOE). Table 1b displays the summary statistics for these variables.

Finally, there are three additional factors. The first two, a time-series momentum factor (Moskowitz, Ooi, and Pedersen, 2012) and a betting-against-beta factor (Frazzini and Pedersen, 2014), come from AQR's data library. See `https://www.aqr.com/library/data-sets` for further details about their construction. The third factor is Pástor and Stambaugh (2003)'s liquidity-risk factor, which is available from L'uboš Pástor's website: `http://faculty.chicagobooth.edu/lubos.pastor/research/liq_data_1962_2013.txt`. See Table 1a for descriptive statistics.

5.3 Estimation Results

If real-world traders face a feature-selection problem, then explicitly modeling this problem using the LASSO should improve out-of-sample return predictability

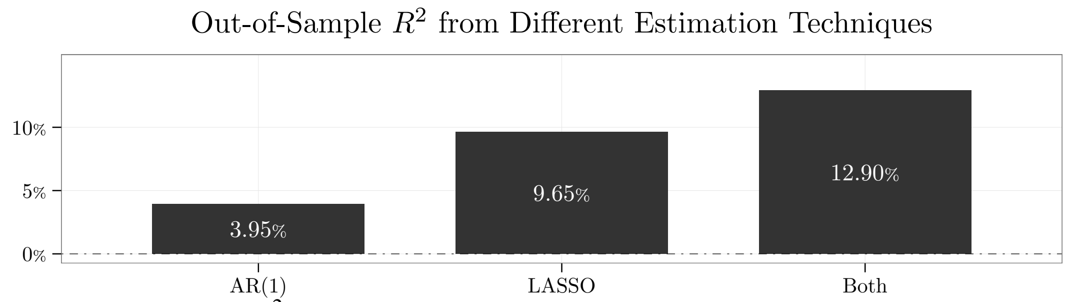


Figure 2: Average R^2 from an out-of-sample prediction of 1-month excess returns using an auto-regression, LASSO, or both. Data: Monthly returns for NYSE-listed stocks from January 1990 to December 2010. Reads: "Using the LASSO to predict returns boosts the out-of-sample R^2 by $(9.65 - 3.95)/3.95 \approx 144.3\%$."

for each stock. By contrast, if traders do not face a feature-selection problem, then the LASSO should not provide any additional out-of-sample predictive power. All the relevant information should be contained in the previous month's returns. The estimation results show that using the LASSO dramatically improves out-of-sample return predictability. Thus, this feature-selection problem appears very important for real-world traders.

Figure 2 houses the key result: using an estimation strategy that accounts for traders' joint inference problem improves the accuracy of out-of-sample predictions. If you just run a simple AR(1) model using rolling 24-month windows, you get an out-of-sample $R^2 = 3.95\%$; whereas, if you estimate the LASSO on the same 24-month windows, you get an out-of-sample $R^2 = 9.65\%$, an improvement of $(9.65 - 3.95)/3.95 \approx 144\%$. Moreover, these two estimation techniques are capturing fundamentally different information. When both the AR(1) and LASSO predictors are included in the same regression, the resulting R^2 is 95% of the sum of the R^2s from the separate regressions.

The pattern of factor loadings that the LASSO chooses also supports the idea that traders face a feature-selection problem. Tables 2a, 2b, 2c, and 2d reveal that, at most, only a small fraction of all NYSE stocks load each of the 79 different factors at any given time. There are many factors which are unhelpful for predicting future stock returns for months on end. For example, Table 2a shows that, while the excess returns to a momentum strategy predict the future returns of 18% of all NYSE stocks in January 2009, momentum is a significant predictor for zero stocks in March 1997.

None of the predictors is a useful indicator for all the stocks all the time. Factors suddenly lurch into importance and then shrink away. It just is not obvious to traders ahead of time which factors they should be using to predict returns. This is exactly the sort of joint inference problem analyzed in the model above.

6 Conclusion

Real-world traders have to simultaneously figure out both which asset features matter and also how much they matter. This paper develops the asset-pricing implications of traders' joint inference problem. Because traders have to simultaneously answer both 'Which features?' and 'How much do they matter?', the risk of selecting the wrong subset of features can spill over, warp their perception of asset values, and distort prices. Thus, feature-selection risk can limit market efficiency even though it stems from the inherent high-dimensional nature of modern asset markets and not some cognitive constraint or trading friction.

References

Ang, A., R. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *Journal of Finance 61*(1), 259–299.

Baker, M. and J. Wurgler (2000). The equity share in new issues and aggregate stock returns. *Journal of Finance 55*(5), 2219–2257.

Baker, M. and J. Wurgler (2004). A catering theory of dividends. *Journal of Finance 59*(3), 1125–1165.

Baker, M. and J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance 61*(4), 1645–1680.

Barberis, N., M. Huang, and T. Santos (2001). Prospect theory and asset prices. *Quarterly Journal of Economics 116*(1), 1–53.

Barberis, N., A. Shleifer, and J. Wurgler (2005). Comovement. *Journal of Financial Economics 75*(2), 283–317.

Black, F. (1986). Noise. *Journal of Finance 41*(3), 529–543.

Campbell, J., M. Lettau, B. Malkiel, and Y. Xu (2001). Have individual stocks become more volatile? an empirical exploration of idiosyncratic risk. *Journal of Finance 56*(1), 1–43.

Candes, E. and Y. Plan (2009). Near-ideal model selection by ℓ_1 minimization. *Annals of Statistics 37*(5), 2145–2177.

Candes, E. and T. Tao (2005). Decoding by linear programming. *IEEE Transactions on Information Theory 51*(12), 4203–4215.

Chen, H., W. Dou, and L. Kogan (2014). Measuring the "dark matter" in asset-pricing models. Working Paper.

Chen, Z. and R. Petkova (2012). Does idiosyncratic volatility proxy for risk exposure? *Review of Financial Studies 25*(9), 2745–2787.

Chinco, A., A. Clark-Joseph, and M. Ye (2019). Sparse signals in the cross-section of returns. *Journal of Finance*.

Cohen, L. and A. Frazzini (2008). Economic links and predictable returns. *Journal of Finance 63*(4), 1977–2011.

Cover, T. and J. Thomas (1991). *Elements of Information Theory* (1 ed.). Wiley Series in Telecommunications.

Daniel, K. (2009). Anatomy of a crisis. *CFA Institute Conference Proceedings Quarterly 26*(3), 11–21.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam (1998). Investor psychology and security market under- and over-reactions. *Journal of Finance 53*(6), 1839–1885.

D'Aspremont, A. and R. Luss (2012). Predicting abnormal returns for news using text classification. *Quantitative Finance 1*(1), 1–12.

DeGroot, M. (1969). *Optimal Statistical Decisions* (1 ed.). McGraw-Hill.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory 52*(4), 1289–1306.

Fama, E. and K. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

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

Gabaix, X. (2013). A sparsity-based model of bounded rationality. *Quarterly Journal of Economics 129*(4), 1661–1710.

Hayek, F. (1945). The use of knowledge in society. *American Economic Review 35*(4), 519–530.

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2014). The common factor in idiosyncratic volatility: Quantitative asset-pricing implications. Working Paper.

Hong, H. and M. Kacperczyk (2009). The price of sin: The effects of social norms on markets. *Journal of Financial Economics 93*(1), 15–36.

Hong, H. and J. Stein (2007). Disagreement and the stock market. *Journal of Economic Perspectives 21*(2), 109–128.

Hong, H., J. Stein, and J. Yu (2007). Simple forecasts and paradigm shifts. *Journal of Finance 62*(3), 1207–1242.

Huberman, G. and T. Regev (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *Journal of Finance 56*(1), 387–396.

Ibbotson, R., J. Sindelar, and J. Ritter (1994). The market's problems with the pricing of initial public offerings. *Journal of Applied Corporate Finance 7*(1), 66–74.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *Journal of Finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock-market efficiency. *Journal of Finance 48*(1), 65–91.

Khandani, A. and A. Lo (2007). What happened to the quants in august 2007? *Journal of Investment Management 5*(1), 29–78.

Klasa, S., W. Maxwell, and H. Ortiz-Molina (2009). The strategic use of corporate cash holdings in collective bargaining with labor unions. *Journal of Financial Economics 92*(3), 421–442.

Kogan, L. and M. Tian (2014). Firm characteristics and empirical factor models: A model-mining experiment. Working Paper.

Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica 53*(6), 1315–1335.

Lo, A. and C. MacKinlay (1990). Data-snooping biases in tests of asset-pricing models. *Review of Financial Studies 3*(3), 431–467.

Miller, E. (1977). Risk, uncertainty, and divergence of opinion. *Journal of Finance 32*(4), 1151–1168.

Moskowitz, T., Y. Ooi, and L. Pedersen (2012). Time-series momentum. *Journal of Financial Economics 104*(2), 2113–2241.

Natarajan, B. (1995). Sparse approximate solutions to linear systems. *SIAM Journal of Computing 24*(2), 227–234.

Neal, R. and S. Wheatley (1998). Do measures of investor sentiment predict returns? *Journal of Financial and Quantitative Analysis 33*(4), 523–547.

Newey, W. and D. McFadden (1994). Large-sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2113–2241.

Pástor, L. and R. Stambaugh (2003). Liquidity risk and expected stock returns. *Journal of Political Economy 111*(3), 642–685.

Rockafellar, T. (1993). Lagrange multipliers and optimality. *SIAM Review 35*(2), 183–238.

Sims, C. (2006). Rational inattention: Beyond the linear-quadratic case. *American Economic Review 96*(2), 158–163.

Sullivan, R., A. Timmermann, and H. White (1999). Data snooping, technical-trading-rule performance, and the bootstrap. *Journal of Finance 54*(5), 1647–1691.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society 58*(1), 267–288.

Wainwright, M. (2009a). Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory 55*(12), 5728–5741.

Wainwright, M. (2009b). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory 55*(5), 2183–2202.

A Proofs

Proof (Proposition 2.3). Each of the N asset-specific informed traders knows his own asset's true value, v_n, and solves the optimization problem below

$$\max_{y_n} \mathbb{E}\left[(v_n - p_n) \cdot y_n \,|\, v_n \right]$$

giving the demand coefficient, $\theta(\lambda)$, up to the determination of λ:

$$y_n = \underbrace{\frac{1}{2 \cdot \lambda}}_{\theta(\lambda)} \cdot v_n$$

I use the notation that $\boldsymbol{X}_{[\mathcal{K}]}$ denotes the measurement matrix $\boldsymbol{X}$ restricted to the columns $\mathcal{K}$ and that $\boldsymbol{\alpha}_{[\mathcal{K}]}$ denotes the coefficient vector $\boldsymbol{\alpha}$ restricted to the elements $\mathcal{K}$. Since an oracle has told the market maker which K features have realized a shock, he can use ordinary least squares to estimate $\boldsymbol{\alpha}$:

$$\hat{\boldsymbol{\alpha}}_{[\mathcal{K}],\text{OLS}} = \left\{ \left(\boldsymbol{X}_{[\mathcal{K}]}^{\top} \boldsymbol{X}_{[\mathcal{K}]} \right)^{-1} \boldsymbol{X}_{[\mathcal{K}]}^{\top} \right\} \frac{\boldsymbol{d}}{\theta(\lambda)}$$

Thus, the cross-section of demand gives the market maker a signal about asset fundamentals

$$\hat{\boldsymbol{v}}_{\text{OLS}} = \boldsymbol{X}_{[\mathcal{K}]} \hat{\boldsymbol{\alpha}}_{[\mathcal{K}],\text{OLS}} = \frac{1}{\theta(\lambda)} \cdot \boldsymbol{d}$$

which has signal error:

$$\mathrm{E}\left[\frac{1}{N} \cdot \|\boldsymbol{v} - \hat{\boldsymbol{v}}_{\text{OLS}}\|_2^2 \right] = \frac{K}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}$$

Least squares prediction errors are normally distributed. In the limit as $N \to \infty$, the asset values are normally distributed since shocks, α_q, are bounded and selected independently from the same distribution. Using DeGroot (1969) updating to compute the market maker's posterior beliefs gives:

$$\mathbb{V}\text{ar}[v_n|\boldsymbol{d}] = \left(\frac{\frac{K}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}}{\frac{K}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2} + \sigma_v^2} \right) \times \sigma_v^2$$

$$\mathbb{E}[v_n|\boldsymbol{d}] = \underbrace{\frac{1}{\theta(\lambda)} \cdot \left(\frac{\sigma_v^2}{\frac{K}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2} + \sigma_v^2} \right)}_{\lambda} \cdot d_n$$

Substituting in $\theta(\lambda) = 1/(2 \cdot \lambda)$ and simplifying gives the desired result. □

Lemma A.1 (Fano's Error Inequality, Cover and Thomas (1991))**.** *Suppose x is a random variable with N outcomes $\{x_1, \ldots, x_N\}$. Let y be a correlated random variable, $\mathbb{C}\mathrm{orr}[x, y] \neq 0$, and let $f(y)$ be the predicted value of x for some deterministic function $f(\cdot)$. Then, the following inequality holds,*

$$\mathbb{P}\mathrm{rob}[x = f(y)] \geq 1 - \frac{\mathrm{M}[x, y]}{\log_2(N)} - \mathrm{o}(1),$$

where $\mathrm{M}[x, y]$ denotes the mutual entropy between the random variables x and y.

Lemma A.2 (Mutual Information Bound, Cover and Thomas (1991))**.** *Suppose p is a random variable with N outcomes $\{p_1, \ldots, p_N\}$ that represent probability distributions of $x \in \mathcal{X}$. Let $\hat{x} \in \mathcal{X}$ be a realization from 1 of the N probability distributions. Then, the following inequality holds,*

$$\mathrm{M}[p, y] \leq \frac{1}{N^2} \cdot \sum_{n,n'=1}^{N} \mathrm{KL}[p_n(x|\hat{x}), p_{n'}(x|\hat{x})],$$

where $\mathrm{KL}[p_n, p_{n'}]$ is the Kullback-Leibler divergence between the distributions p_n and $p_{n'}$.

Proof (Proposition 3.1). I show that if there exists some fixed constant C such that

$$N < C \cdot K_N \cdot \log(Q_N / K_N)$$

as $N \to \infty$, then there does not exist an inference rule $\phi \in \mathbf{\Phi}$ such that $\mathrm{FSE}[\phi] \to 0$.

The proof proceeds in 6 steps:

1. *Define variables.* Let $S = \binom{Q}{K}$ denote the number of feature subsets of size K and index each of these subsets with $\mathcal{K}_s$ for $s = 1, 2, \ldots, S$. It is sufficient to consider the case where $\alpha_q = \alpha_{\min}$ for all $q \in \mathcal{K}_\star$ since this is easiest case. If there is no selection rule ϕ that can identify the correct subset $\mathcal{K}$ when all of the coefficients are fixed at $\alpha_{\min}$, then there can be none when the coefficients are variable. Each subset is then associated with a distribution, p_s, given by

$$p_s = \mathrm{Normal}(\alpha_{\min} \cdot \boldsymbol{X}[\mathcal{K}_s]\mathbf{1}, \boldsymbol{I})$$

for $s = 1,2,\ldots,S$ where $\boldsymbol{X}[\mathcal{K}_s]$ denotes the observed measurement matrix restricted to the columns $\mathcal{K}_s$, $\mathbf{1}$ denotes a $(K \times 1)$-dimensional vector of 1s, and $\boldsymbol{I}$ denotes the $(K \times K)$-dimensional identity matrix.

2. *Apply information inequalities.* Picking the right subset, $s \in \{1, \ldots, S\}$, then amounts to picking the right generating distribution. Fano's inequality says that

$$\text{FSE}[\phi] = \mathbb{P}\text{rob}[\mathcal{K} = \phi(\boldsymbol{d}, \boldsymbol{X})] \geq 1 - \frac{\text{M}[p, \boldsymbol{d}|\boldsymbol{X}]}{\log_2(S)} - \text{o}(1)$$

I want to find conditions under which the right-hand side of this inequality is greater than 0. To do this, I need to characterize $\text{M}[p, \boldsymbol{d}|\boldsymbol{X}]$ which can be upper bounded:

$$\text{M}[p, \boldsymbol{d}|\boldsymbol{X}] \leq \frac{1}{S^2} \cdot \sum_{s,s'=1}^{S} \text{KL}[p_s(\boldsymbol{d}'|\boldsymbol{d}, \boldsymbol{X}), p_{s'}(\boldsymbol{d}'|\boldsymbol{d}, \boldsymbol{X})]$$

3. *Use functional form.* The optimal selection rule searches over all S feature subsets and tries to solve the program

$$\min_{s=1,2,\ldots,S} \|\boldsymbol{d} - \alpha_{\min} \cdot \boldsymbol{X}[\mathcal{K}_s]\mathbf{1}\|_2^2 = \min_{s=1,2,\ldots,S} \|\alpha_{\min} \cdot (\boldsymbol{X}[\mathcal{K}_\star] - \boldsymbol{X}[\mathcal{K}_s])\mathbf{1} + \boldsymbol{\epsilon}\|_2^2$$

Plugging in the form of the optimization problem to characterize the Kullback-Leibler divergence and rearranging then gives

$$\begin{aligned} \text{FSE}[\phi] &= \mathbb{P}\text{rob}[\mathcal{K} = \phi(\boldsymbol{d}, \boldsymbol{X})] \\ &\geq 1 - \left(\frac{\frac{1}{2 \cdot S^2} \cdot \sum_{s,s'=1}^{S} \|\alpha_{\min} \cdot (\boldsymbol{X}[\mathcal{K}_s] - \boldsymbol{X}[\mathcal{K}_{s'}])\mathbf{1}\|_2^2}{\log_2(S)} \right) - \text{o}(1) \end{aligned}$$

In order for $\text{FSE}[\phi] > 0$, it has to be the case that as $N \to \infty$ we have that

$$1 > \frac{1}{2 \cdot S^2} \cdot \frac{\sum_{s,s'=1}^{S} \|\alpha_{\min} \cdot (\boldsymbol{X}[\mathcal{K}_s] - \boldsymbol{X}[\mathcal{K}_{s'}])\mathbf{1}\|_2^2}{\log_2(S)}$$

4. *Characterize error distribution.* For any pair of subsets $(\mathcal{K}_s, \mathcal{K}_{s'})$ define the random variable as

$$h_{s,s'} = \|\alpha_{\min} \cdot (\boldsymbol{X}[\mathcal{K}_s] - \boldsymbol{X}[\mathcal{K}_{s'}])\,\mathbf{1}\|_2^2$$

Because assets have feature exposures, $x_{n,q} \overset{\text{iid}}{\sim} \text{Normal}(0,1)$, $h_{s,s'}$ follows a χ_N^2 distribution,

$$h_{s,s'} \sim 2 \cdot \alpha_{\min}^2 \cdot (K - |\mathcal{K}_s \cap \mathcal{K}_{s'}|) \cdot \chi_N^2$$

where $|\mathcal{K}_s \cap \mathcal{K}_{s'}|$ denotes the size of the set difference between the subsets $\mathcal{K}_s$ and $\mathcal{K}_{s'}$. For example, if there are $K = 4$ shocked features and $\mathcal{K}_s = \{1, 2, 5, 9\}$

while $\mathcal{K}_{s'} = \{1, 3, 5, 9\}$, then $|\mathcal{K}_s \cap \mathcal{K}_{s'}| = 1$.

5. *Bound mass in tail.* Using the tail bound for a χ^2_N distribution, we see that

$$\mathbb{P}\text{rob}\left[\frac{1}{S^2} \cdot \sum_{s \neq s'} h_{s,s'} \geq 4 \cdot \alpha^2_{\min} \cdot K \cdot N\right] \leq 1/2$$

Thus, at least half of the S different subsets obey the bound below:

$$\frac{1}{2 \cdot S^2} \cdot \frac{\sum_{s,s'=1}^{S} \|\alpha_{\min} \cdot (\boldsymbol{X}[\mathcal{K}_s] - \boldsymbol{X}[\mathcal{K}_{s'}])\mathbf{1}\|_2^2}{\log(S)} \leq \frac{4 \cdot \alpha^2_{\min} \cdot K \cdot N}{\log_2(S)}$$

6. *Formulate key inequality.* Thus, as long as

$$1 > \frac{4 \cdot \alpha^2_{\min} \cdot K \cdot N}{\log_2(S)}$$

holds, the error rate will remain bounded away from 0 implying that

$$N > \left(\frac{1}{4 \cdot \alpha^2_{\min} \cdot K}\right) \times \log_2(S)$$

is necessary for $\text{FSE}[\phi] \to 0$. The multiplier $(4 \cdot \alpha^2_{\min} \cdot K)^{-1}$ is where the fixed constant C comes from in the result, so it is obvious that the constant will depend on the way that $\alpha_{\min}$ and K scale as the market grows large. To make the formula above match, simply recall that:

$$S = \binom{Q}{K} \geq \left(\frac{Q}{K}\right)^K$$

□

Lemma A.3 (Bound on Signal Error, Candes and Plan (2009)). *If $N \geq N^\star(Q, K)$, then the LASSO estimate, $\hat{\boldsymbol{\alpha}}_{LASSO}$, from the program in equation (20) using the tuning parameter $\gamma = 2 \cdot (\sigma_z/\theta) \cdot \sqrt{2 \cdot \log(Q)}$ obeys the inequality,*

$$\mathbb{P}\text{rob}\left[\frac{1}{N} \cdot \|\boldsymbol{X\alpha} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\|_2^2 \leq \tilde{C}^2 \times \left(\frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta^2}\right)\right] \geq 1 - \frac{6}{Q^{2 \cdot \log 2}} - \frac{1}{Q \cdot \sqrt{2 \cdot \pi \cdot \log(Q)}}$$

with numerical constant $\tilde{C} = 4 \cdot (1 + \sqrt{2})$.

Proof (Proposition 4.2). Just as in Proposition 2.3, each of the N asset-specific informed traders knows his own asset's true value, v_n, and solves

$$\max_{y_n} \mathbb{E}\left[\,(v_n - p_n) \cdot y_n \,|\, v_n\,\right]$$

giving the demand coefficient, $\theta(\lambda)$, up to the determination of λ:

$$y_n = v_n/(2 \cdot \lambda) = v_n \cdot \theta(\lambda)$$

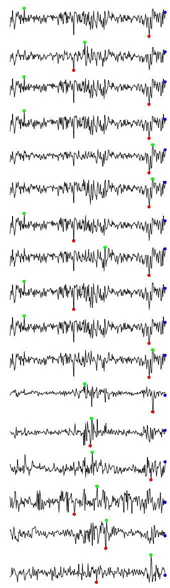
In the limit as $N \to \infty$, the asset values are normally distributed since shocks, α_q, are bounded and selected independently from the same distribution. However, now the cross-section of aggregate demand gives a signal about each asset's fundamental value with mean $\boldsymbol{v}$ and variance given in Lemma A.3.

Using DeGroot (1969) updating to compute the market maker's posterior beliefs gives:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[v_n|\boldsymbol{d}] &= \left(\frac{C^2 \cdot \frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}}{\sigma_v^2 + C^2 \cdot \frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}} \right) \times \sigma_v^2 \\
\mathbb{E}[v_n|\boldsymbol{d}] &= \underbrace{\frac{1}{\theta} \cdot \left(\frac{\sigma_v^2}{\sigma_v^2 + C^2 \cdot \frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}} \right)}_{\lambda} \cdot \boldsymbol{d}
\end{aligned}$$

Noting that $\theta(\lambda) = 1/(2 \cdot \lambda)$ then gives the desired result after simplifying. □

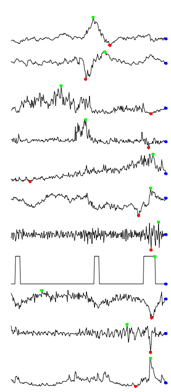
Summary Statistics for Market Factors



	Avg	Med	StD	Min	Max
Value-Weighted Market	0.5	1.1	4.5	−17.2	10.8
Small Stocks	1.1	1.7	6.2	−21.5	24.2
Large Stocks	0.8	1.2	4.4	−16.4	11.6
Growth Stocks	0.8	1.1	4.6	−15.4	14.2
Value Stocks	0.9	1.6	4.8	−22.1	16.6
Low Op. Profit Stocks	0.6	1.4	5.5	−21.4	12.5
High Op. Profit Stocks	0.9	1.3	4.2	−15.4	13.2
Low Investment Stocks	0.9	1.5	4.3	−16.1	10.5
High Investment Stocks	0.8	1.3	5.4	−18.5	13.6
Low E/P Stocks	0.8	1.1	4.7	−16.3	13.4
High E/P Stocks	1.1	1.6	4.6	−18.5	12.5
Momentum	0.6	0.8	5.3	−34.7	18.4
Short-Term Reversal	0.3	0.2	3.8	−14.5	16.2
Long-Term Reversal	0.4	0.3	2.6	−7.0	11.1
Time-Series Momentum	1.7	1.5	7.9	−17.4	24.4
Betting Against Beta	0.8	1.0	3.2	−10.5	10.7
Liquidity Factor	0.6	0.4	4.0	−10.1	21.0

Table 1a: Monthly excess return on market factors from January 1990 to December 2010. Sources: Ken French's data library, AQR's data library, and L'uboš Pástor's website.

Summary Statistics for Sentiment and Macro Variables



	Avg	Med	StD	Min	Max
Sentiment Index	0.1	0	0.5	−0.9	2.5
Dividend Premium	−6.9	−7.2	9.8	−50.2	17.1
Number of IPOs	27.4	21.5	22	0	106
Return on IPOs	17.9	13.7	19.6	−19.9	116.2
Turnover	0.8	0.8	0.3	0.4	1.7
Closed-End-Fund Discount	6.1	6.2	4.5	−6	18.2
Equity Share in New Issues	0.3	2.1	57.9	−246.3	204.6
Recession Indicator	0.1	0	0.3	0	1
Employment Growth	0.1	0.1	0.2	−0.6	0.4
Inflation Rate	0.2	0.2	0.3	−1.9	1.2
VIX	20.4	19.2	8	10.8	62.6

Table 1b: Monthly values for a variety of sentiment and macroeconomic variables from January 1990 to December 2010. Source: Jeffrey Wurgler's website, NBER, BEA, and CBOE.

Summary Statistics for Country-Specific Factors

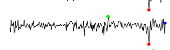
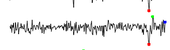
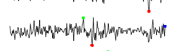
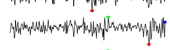
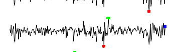
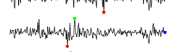
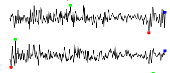
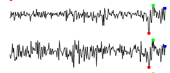
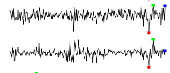
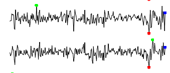
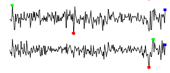
		Avg	Med	StD	Min	Max
Austria		0.8	1.0	7.0	−34.4	20.0
Australia		1.1	1.0	6.0	−27.7	17.3
Belgium		0.8	1.0	5.6	−30.8	16.7
Canada		1.0	1.4	5.6	−26.9	21.5
Denmark		1.0	1.4	5.7	−25.6	18.2
Finland		1.3	1.0	9.1	−29.0	30.8
France		0.8	1.2	5.9	−21.6	15.0
Germany		0.8	1.2	6.3	−23.3	22.0
Hongkong		1.4	1.2	7.6	−28.6	32.2
Italy		0.6	0.8	7.2	−23.7	21.2
Japan		0.1	0.0	6.4	−18.6	25.9
Netherlands		1.0	1.3	5.9	−30.1	17.1
New Zealand		0.6	0.9	6.4	−19.0	15.6
Norway		1.1	1.3	7.5	−31.0	19.5
Singapore		1.0	1.3	7.4	−28.3	29.3
Spain		0.9	1.0	6.7	−22.8	21.1
Sweden		1.2	1.5	7.6	−27.6	25.2
Switzerland		1.0	1.3	4.8	−14.7	14.8
United Kingdom		0.8	0.8	4.9	−20.4	14.3

Table 1c: Monthly excess return on value-weighted country-specific portfolios from January 1990 to December 2010. Source: Ken French's data library.

Summary Statistics for Industry-Specific Factors

	Avg	Med	StD	Min	Max
Automobiles	0.8	1.0	8.0	−36.4	49.6
Beer	1.0	1.3	5.2	−19.8	16.4
Books	0.5	0.3	5.7	−26.5	33.1
Business Equipment	1.1	1.3	8.3	−31.8	25.1
Chemicals	0.9	1.2	5.7	−20.9	22.1
Clothes	1.0	1.6	6.7	−22.1	25.1
Coal	2.0	1.4	11.8	−37.9	44.0
Construction	0.8	1.2	6.1	−28.2	23.3
Electrical Equipment	1.3	1.1	6.5	−24.6	23.2
Fabricated Products	1.0	1.5	6.8	−29.9	20.8
Financials	0.9	1.3	5.8	−22.1	17.0
Food	0.9	1.1	4.2	−12.1	15.7
Gaming	1.0	1.3	7.1	−29.7	34.5
Healthcare	0.9	1.1	4.6	−12.3	16.5
Household	0.9	1.2	4.5	−14.3	18.5
Non-Auto Vehicles	1.2	1.8	6.2	−24.1	17.1
Non-Coal Mining	1.0	1.2	8.1	−34.5	35.6
Oil and Gas	1.0	0.7	5.3	−16.9	19.1
Other	0.3	0.6	6.0	−21.3	19.8
Paper	0.8	1.0	5.1	−18.5	21.0
Services	1.1	1.8	6.9	−19.3	23.8
Restaurants	0.9	1.3	5.2	−14.8	16.0
Retail	0.9	0.9	5.3	−14.6	14.3
Steel	0.9	0.9	8.6	−33.0	30.7
Tabacco	1.2	2.0	7.2	−24.9	32.5
Telecom	0.6	1.2	5.5	−16.2	21.3
Textiles	0.6	1.0	8.6	−28.5	59.0
Transportation	0.9	1.4	5.3	−16.7	14.5
Utilities	0.8	1.2	4.2	−12.7	11.7
Wholesale	0.7	1.3	4.8	−21.1	15.2

Table 1d: Monthly excess return on value-weighted industry-specific portfolios from January 1990 to December 2010. Source: Ken French's data library.

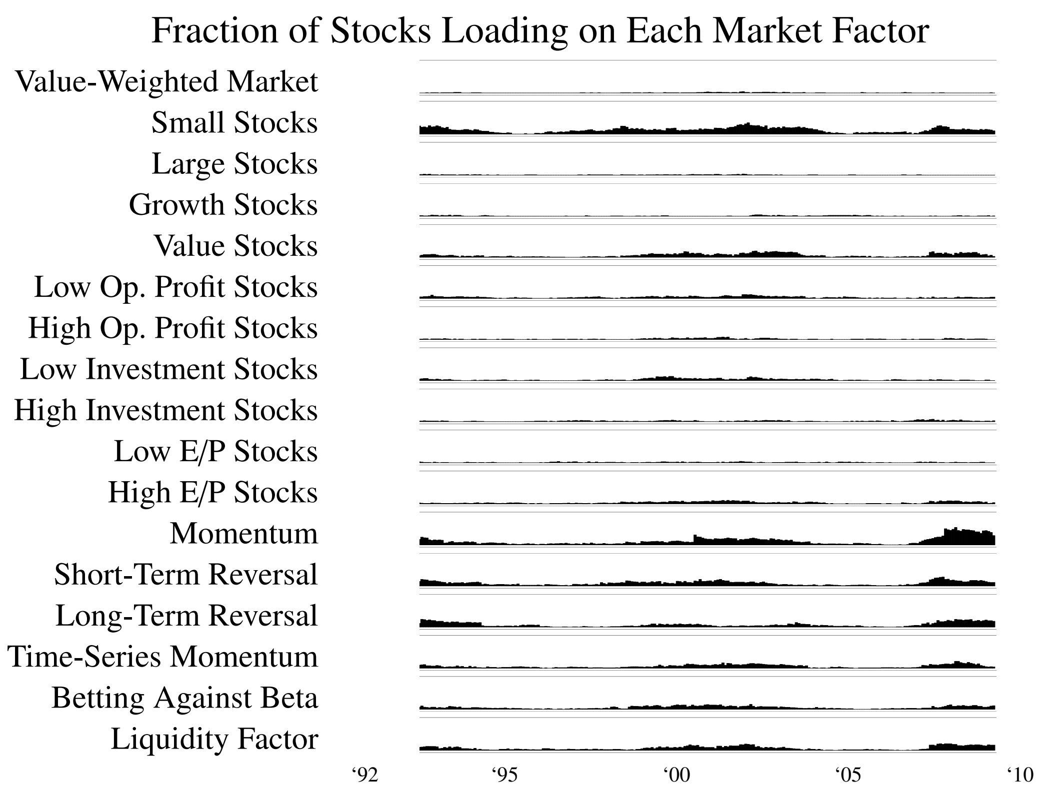


Table 2a: Fraction of NYSE stocks each month that have non-zero loadings on each market factor when estimated using the LASSO in Equation (28). y-axis ranges from 0% to 25%.

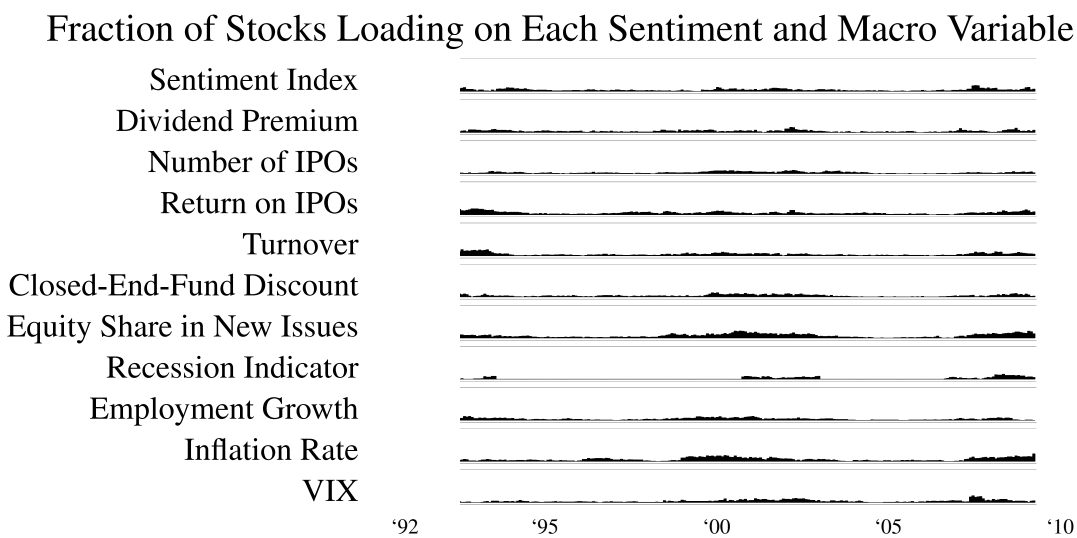


Table 2b: Fraction of NYSE stocks each month that have non-zero loadings on each sentiment and macroeconomic variarable when estimated using the LASSO in Equation (28). *y*-axis ranges from 0% to 25%.

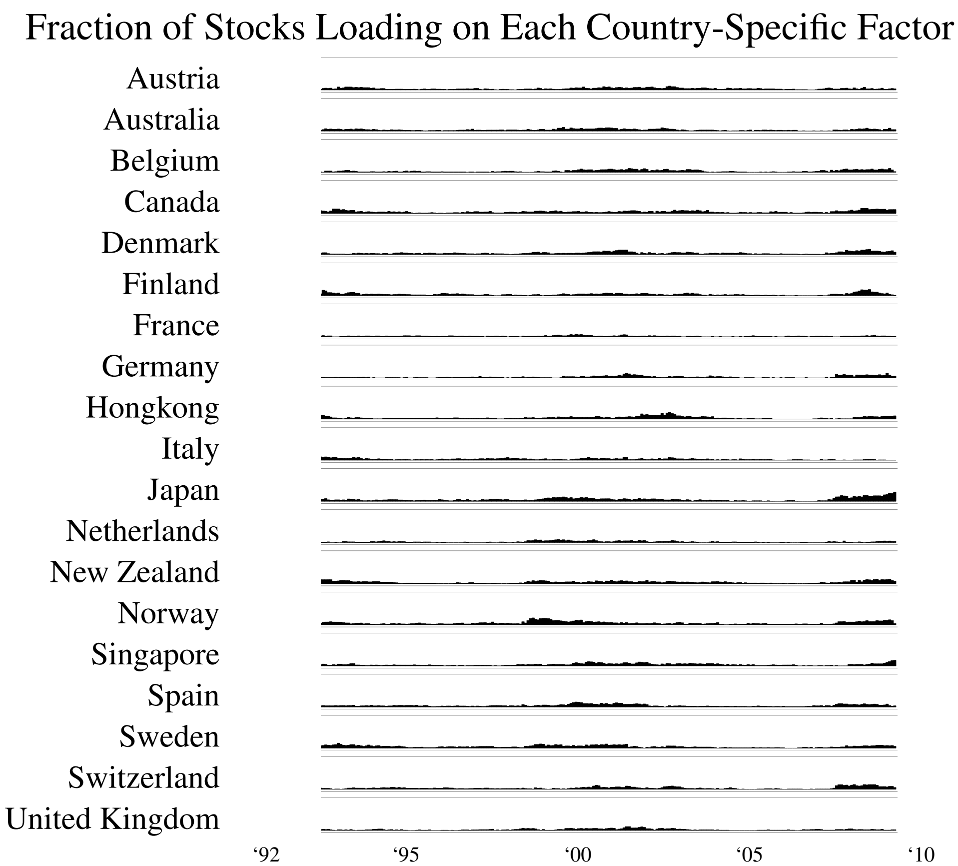


Table 2c: Fraction of NYSE stocks each month that have non-zero loadings on each country-specific factor when estimated using the LASSO in Equation (28). y-axis ranges from 0% to 25%.

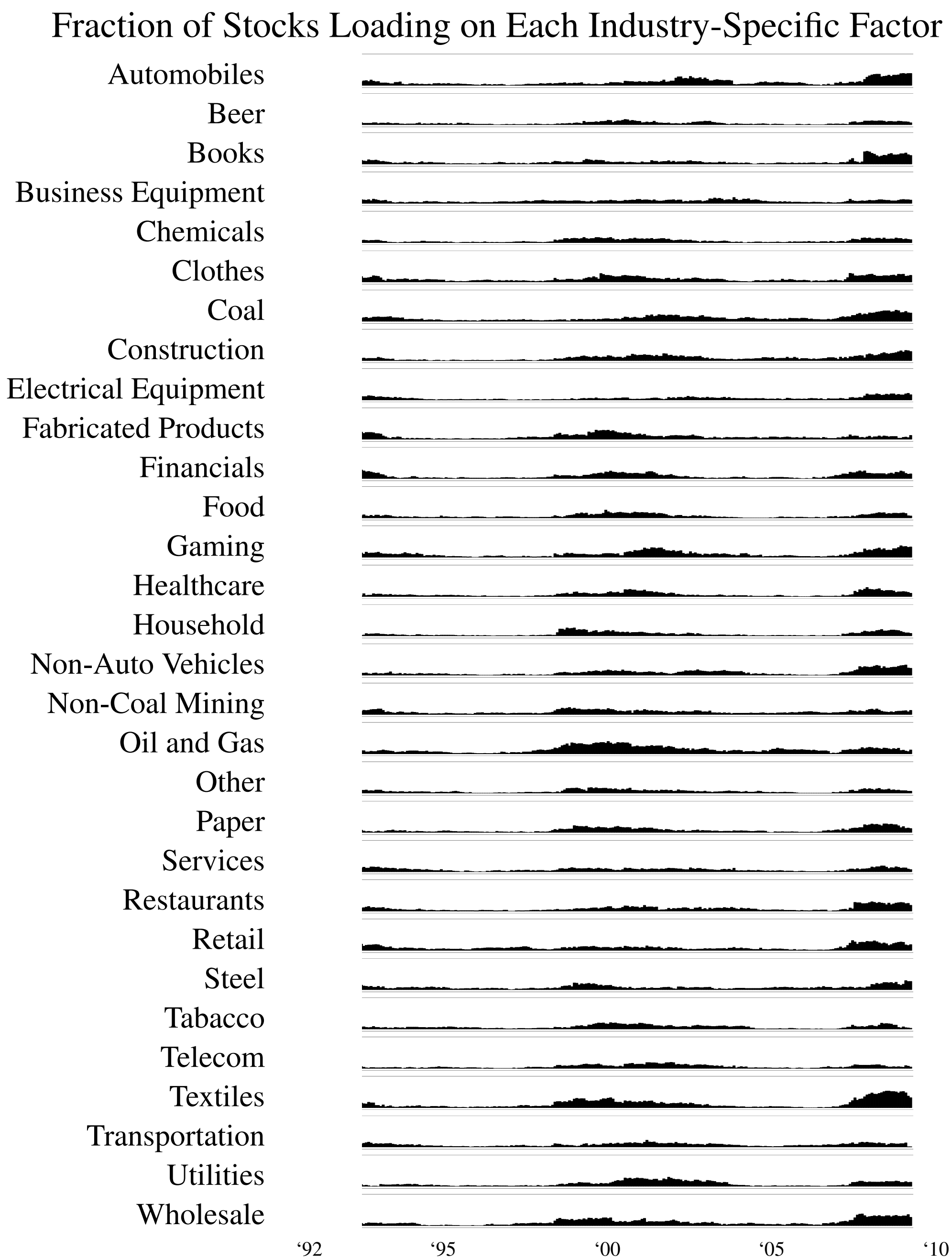


Table 2d: Fraction of NYSE stocks each month that have non-zero loadings on each industry-specific factor when estimated using the LASSO in Equation (28). y-axis ranges from 0% to 25%.