FEATURE-SELECTION RISK

ALEX CHINCO

Abstract. Companies have exposure to many different features that might plausibly affect their stock returns, like whether they're involved in a crowded trade, whether they're mentioned in an M&A rumor, or whether their supplier missed an earnings forecast. Yet, at any point in time, only a handful of these features actually matter. As a result, real-world traders have to simultaneously infer both the identity and the value of the few relevant features. This paper shows that, because they face this joint inference problem, the risk of selecting the wrong subset of features can spill over and warp traders' perceived asset values. The high-dimensional nature of modern financial markets can act as a limit to arbitrage.

JEL Classification. D83, G02, G12, G14

Keywords. Feature-Selection Risk, Limits to Arbitrage, Sparsity, Behavioral Finance

Date: November 28, 2014.

University of Illinois at Urbana-Champaign. alexchinco@gmail.com. (916) 709-9934.

I am extremely indebted to Xavier Gabaix for many extremely enlightening conversations about this topic. I have also received many helpful comments and suggestions from Adam Clark-Joseph, Aurel Hizmo, Ron Kaniel, Vuk Talijan, and Jeff Wurgler as well as seminar participants at the Academy of Behavioral Finance Conference, UIUC (Finance), and Rochester (Simon).

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

1. Introduction

Real-world traders have to simultaneously figure out both which asset features matter and also how much they matter. You can find evidence of this joint inference problem throughout modern financial markets. To begin with, quant-fund pitch books are studded with phrases like, "our model allows us to identify and interpret events faster than more traditional methods used by other investors."[1] Alternatively, notice how trading floors are covered in row after row of multi-monitor displays. These hi-tech orchards wouldn't exist in a world where traders already knew which features to analyze. You can even hear the problem echoed back in traders' war stories. "Before the 1998 financial crisis began, I didn't even know who LTCM was," recalls Colm O'Shea, founder of COMAC Capital.[2] "At the start of the crisis, there was nothing about LTCM in the press... All I knew was that T-bond futures were going up limit every day. That told me there was something going on."

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 act like a limit to arbitrage even though it stems from the inherent high-dimensional nature of modern asset markets and not some cognitive constraint or trading friction.

Illustrative Example. Let's take a look at a short example illustrating exactly why feature-selection risk is a consequence of a market's dimensions and exactly how it limits arbitrage. Imagine you're a trader in a market where each company's stock returns can have exposure to any combination of 7 features: 1) whether it's involved in a crowded trade (Khandani and Lo (2007)), 2) whether it's been mentioned in a news article about M&A activity (D'Aspremont and Luss (2012)), 3) whether there's been an announcement about its major supplier (Cohen and Frazzini (2008)), 4) whether its labor force has unionized (Klasa, Maxwell, and Ortiz-Molina (2009)), 5) whether it belongs to the alcohol, tobacco, and gaming industry (Hong and Kacperczyk (2009)), 6) whether it's been referenced in a scientific journal article (Huberman and Regev (2001)), and 7) whether it's been included in the S&P 500 (Barberis, Shleifer, and Wurgler (2005)).

Moreover, suppose you know that companies with 1 of these 7 features might have realized a shock, but you don't initially know which one. All you know is that the market hasn't fully appreciated the shock. Stocks with this mystery feature will realize abnormal returns of $\alpha > 0$ percent which for the next few trading periods. Here is the question. How many stock returns do you need to see in order to figure out which, if any, of the shocks has occurred?

Answer: 3.

[1] Actual quote from pitch book of quantitative trading desk at fund with more than $1 trillion AUM.
[2] Schwager, J. (1992) Market Wizards (1 ed.) John Wiley & Sons.

Three observations give just enough information to answer 7 yes-or-no questions and rule out the possibility of no change, $7 = 2^3 - 1$. Let's construct the solution to see why. Suppose the first company's returns have exposure to features {1,3,5,7}—that is, it's involved in a crowded trade, there's been an announcement about its major supplier, it belongs to the alcohol, tobacco, and gaming (ATG) industry, and it's been recently added to the S&P 500. Similarly, suppose that the second company's returns have exposure to features {2,3,6,7} and the third company's returns have exposure to features {4,5,6,7}. The abnormal returns for these three stocks always reveal exactly which feature-specific shock has occurred. If only the first stock has positive abnormal returns, $ar_1 = \alpha$ while $ar_2 = ar_3 = 0$, then there must have been a crowded-trade-specific shock:

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

Whereas, if both the first and the third stock have positive abnormal returns, $ar_1 = ar_3 = \alpha$ while $ar_2 = 0$, then there must have been a shock to the ATG industry.

Now, let's rewind the clock a bit and consider the problem you face after seeing only the first two company's abnormal returns, $ar_1 = \alpha$ while $ar_2 = 0$:

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

In this setting, you know that either a crowded-trade-specific shock has occurred or an ATG-industry-specific shock has occurred since the first company's returns have exposure to both these features while the second company's returns don't—that is, both the first and the fifth columns are $[1\ 0]^\top$. What's the right way to value the third company's stock which has exposure to features {4,5,6,7}, meaning that it's in the ATG industry but it's not involved in a crowded trade?

There are two possibilities. If the crowded-trade-specific shock has occurred, then you should leave the third company's value unchanged; whereas, if the ATG-industry-specific shock has occurred, then you should revise your valuation of the third company's stock. Thus, after seeing only two observations, you have to split the difference. If it was, in fact, the ATG-industry-specific shock, then you will only update the third company's value half-way. So, it will look like you were slow to react to public information. By contrast, if it was the crowded-trade-specific shock, then, when you revise your valuation of the third company's stock half-way, it will look like you were trading on noise. Nevertheless, you

were doing the best you could in real time. It's not like you're making some cognitive error or fighting against some trading friction. Instead, it's the dimensionality of your inference problem that generates the extra risk, that warps your perception of the third company's value, that distorts the third company's stock price.

Feature-Selection Bound. 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 don't reflect standard risk factors, like the return on the market portfolio. To address these concerns, I apply results from the compressed-sensing literature to generalize this thresholding result. Specifically, I show that prior to seeing $N^\star(Q,K)$ observations,[3]

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

it is impossible to consistently 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 each company's feature exposures are randomly assigned and their fundamental values reflect the standard risk factors. What's more, in the presence of noise, some feature-selection risk will remain even after the feature-selection bound has been reached. Thus, the feature-selection bound is an existence result. It says that, in any large market, traders will face some feature-selection risk no matter what optimization program they might use since the risk stems from the dimensionality of the market and not cognitive constraints or trading frictions.

Asset-Pricing Model. Having shown that feature-selection risk is endemic to any large market, I next investigate how it warps traders' perception of asset values, distorts prices, and delays arbitrage. To do this, I study a portable extension of the static Kyle (1985) model with N assets whose values are a function of $K \ll Q$ feature-specific shocks. The equilibrium concept is completely standard. There are many informed traders who observe private signals about the value of a single asset and then submit market orders to a common market maker. And, just like in the original model, competitive pressures force the market maker to set the price of each asset as close as possible to its fundamental value after observing the combined demand from informed and noise traders. The model gets interesting when you ask: 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?

The feature-selection bound implies that there are two regimes. When there are sufficiently few stocks relative to how complex the market is—that is, when $N < N^\star(Q,K)$—the answer is: nothing. Recall that, in the earlier example when you'd only seen the first two assets, you could still tell that the first company had realized a feature-specific shock while the second company hadn't. You just couldn't tell which feature-specific shock. It could've

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

been either the crowded-trade shock or the ATG-industry shock. Similarly, in the model, when $N < N^\star(Q,K)$ each asset's aggregate demand reveals something about that particular asset's fundamental value. The market maker just can't tell which feature-specific shocks are responsible. In the general case, it could be any combination of the $\binom{Q}{K} \gg N$ possibilities with equal probability. Thus, when the market is sufficiently complex relative to the number of assets, it effectively collapses to the original Kyle (1985) equilibrium with the market maker setting prices on an asset-by-asset basis.

By contrast, when the number of stocks crosses the feature-selection bound—that is, when $N \geq N^\star(Q,K)$—the market maker can suddenly learn something about which feature-specific shocks have occurred from the cross-section of aggregate demand. He can use the underlying shock structure to form more accurate beliefs about each asset's fundamental value. Yet, because he now has to infer both which K features matter and also how much they matter, the market maker will be less responsive to aggregate demand shocks and prices will be less accurate relative to the original Kyle (1985) equilibrium. A pair of stocks with exposure to the same feature-specific shock might have different prices because the market maker (and any would-be arbitrageur) doesn't know ahead of time how to interpret the pair of demand shocks. Was their demand saying something about a shared feature? If so, which one? Or, did both these stocks just happen to realize positive noise shocks at the same time?

Empirical Predictions. This model makes a pair of novel empirical predictions. First, the model predicts that assets with more features should have lower pricing-impact coefficients. A market maker will be more likely to make a feature selection error if he has to sort through a larger number of potentially relevant features. So, because he's aware that he's more error prone, he will be less responsive to informed trader demand. Consistent with this prediction, quantitative traders often look for signal which are as obscure as possible. They try to hide in the strategy space with the largest ambient dimension, Q. For example, the co-CEO of Renaissance Technologies, Robert Mercer, 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."[4]

Second, the model makes a prediction about the portfolio of assets that sophisticated arbitrageurs should trade, about the kind of assets that are collectively the most informative about feature-specific shocks. One Arrow security for each risk. This is the textbook approach to learning about risks. However, compressed-sensing theory says that an astute trader can identify feature-specific shocks using far fewer assets if the shocks are sparse and the assets are extremely complicated and heterogeneous. So, to identify feature-specific shocks using as few assets as possible, sophisticated arbitrageurs should simultaneously trade a diverse collection of complex derivatives rather than simpler assets like stocks or bonds.

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

2. Baseline Equilibrium Model

I begin by characterizing a baseline equilibrium where traders don't face any feature-selection risk. Specifically, I assume that they have access to an oracle that alerts them to the K features that have realized a shock, but not the size or sign of the shock. We can then return to this model during the later analysis as a point of comparison to answer the question, 'How does feature-selection risk alter the usual predictions?'.

2.1. Market Structure.

I study a static market with N assets whose fundamental values, v_n, are governed by their exposure to Q features:

$$v_n = \sum_q \alpha_q \cdot x_{n,q} \tag{4}$$

where $x_{n,q} \overset{\text{iid}}{\sim} \text{N}(0,1)$ 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 $\alpha_{ATGind} = \$1$ to stocks in the alcohol, tobacco, and gaming industry, then the share price of a company in that industry, $x_{n,ATGind} = 1$, will rise by \$1.

Heterogeneous 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 opposite affects on the acquirer and the target, and some companies are more strongly impacted by news about a particular supplier than others. Let's consider a short concrete example with only 2 assets. Suppose that asset 1 has exposures to the stat-arb-strategy, M&A activity, and supplier stock features given by $\mathbf{x}_1 = [\,1.50\ 0.50\ -0.10\,]^\top$ while asset 2 has feature exposures $\mathbf{x}_2 = [\,-0.50\ -0.75\ 1.00\,]^\top$. 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.

What's more, I consider a setting where everyone knows each asset's feature exposures $\mathbf{x}_n$—that is, all agents have a detailed list of whether or not each asset's 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 alcohol, tobacco, and gaming industry has realized a shock, but they will never be uncertain about whether a particular company is in the industry.

Sparse Shocks. Only K of the elements in $\boldsymbol{\alpha}$ are non-zero:

$$K = \|\boldsymbol{\alpha}\|_0 = \sum_q 1_{\{\alpha_q \neq 0\}} \qquad \text{with} \qquad Q \gg N \geq K \tag{6}$$

I assume the vector of feature-specific shocks, $\boldsymbol{\alpha}$, satisfies:

(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 $\alpha_{\max} \geq |\alpha_q| > \sigma_z$.

These restrictions on $\boldsymbol{\alpha}$ capture the idea that only a few of the many possible features that might impact a stock's value each period actually matter. Traders have to figure out which ones to pay attention to in real time. 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 larger than those to the remaining $(Q-K)$ features, $K \leq \sum_q 1_{\{|\alpha_q|>\delta\}}$.

Feature Selection. This market structure means that it's 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."[5] This wasn't 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 aren't in the business of spotting feature-specific shocks. For example, a January 2008 Chicago Tribune article about Priceline.com (PCLN) 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."[6] No 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 market-wide market makers, along with a collection of asset-specific noise traders.

Informed Traders' Problem. Asset-specific informed traders know the fundamental value of a single asset, v_n, and solve the standard static Kyle (1985)-type optimization problem

[5]Schwager, J. (1989) *Market Wizards: Interviews with Top Traders.* (1 ed.) New York Institute of Finance.
[6]DiColo, J. (2008, Jan. 20) Priceline's Power Looks Promising in Europe, Asia. *Chicago Tribune.*

with risk neutral preferences,

$$\max_{y_n \in \mathbf{R}} \mathrm{E}\left[\,(v_n - p_n) \cdot y_n \,|\, v_n\,\right] \tag{7}$$

where y_n denotes the size of asset n's informed trader's market order in units of shares. Crucially, for these traders, the fundamental value of each asset is just a random variable with no further structure. They cannot observe which K feature-specific shocks govern its value. It's productive to 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?"[7] It's like they have an informative gut instinct.

Market Maker's Problem. The market-wide market maker observes aggregate order flow, d_n, for each of the N assets,

$$d_n = y_n + z_n \qquad \text{with} \qquad z_n \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma_z^2) \tag{8}$$

which is composed of demand from the asset-specific informed traders, y_n, and from asset-specific noise traders, z_n. He then tries to set the price of each asset as close as possible to its fundamental value given the cross-section of aggregate demand:

$$\min_{\mathbf{p} \in \mathbf{R}^N} \mathrm{E}\left(\frac{1}{N} \cdot \sum_n (p_n - v_n)^2 \,\middle|\, \mathbf{d} \right) \tag{9}$$

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's 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 $\phi(\mathbf{d},\mathbf{X})$ that accepts an $(N \times 1)$-dimensional vector of aggregate demand as well as an $(N \times Q)$-dimensional matrix of features and spits out a vector of feature-specific shocks:

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

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

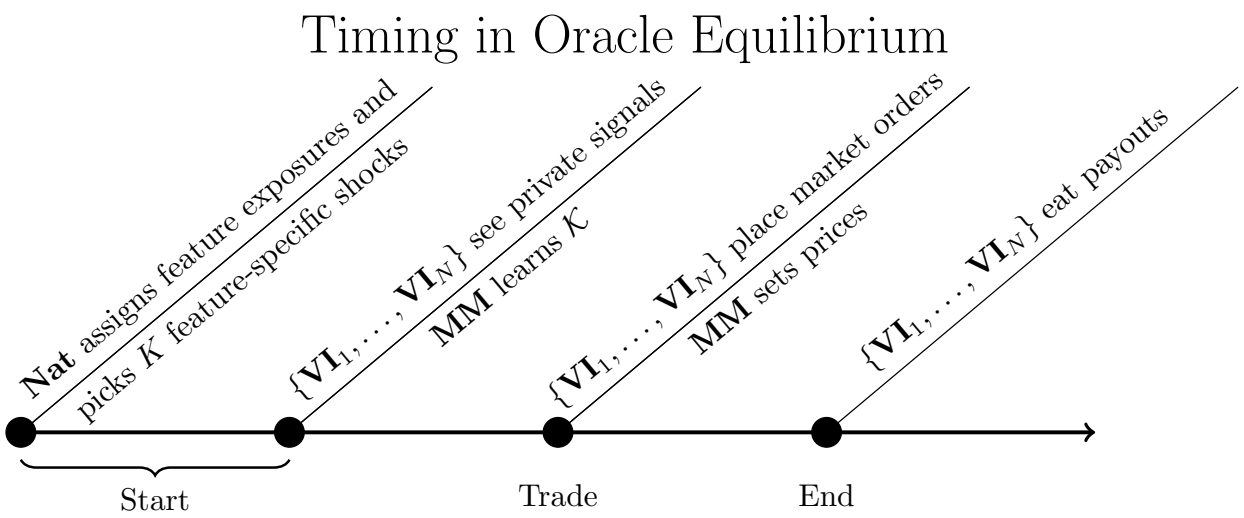

Figure 1. *What each agent knows and when they know it in the model where the common market maker knows which K features have realized a shock.*

I use $\widehat{\boldsymbol{\alpha}} = \phi(\mathbf{d},\mathbf{X})$ to denote 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, $\mathbf{d}$. It's clear that real world traders worry about how much their market maker can learn from the *combination* of their orders. For instance, quantitative hedge funds place the orders for different legs of the same trade with different brokers to make it difficult for their brokers to do exactly 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 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, $\mathcal{E} = \{\theta,\lambda\}$, is a linear demand rule for each of the N asset-specific informed

traders:

$$y_n = \theta \cdot v_n \tag{11}$$

and a linear pricing rule for the single market maker common to all N assets:

$$p_n = \lambda \cdot d_n \tag{12}$$

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:*

$$\lambda = \frac{1}{2 \cdot \theta} \tag{13a}$$

$$\theta = \sqrt{\frac{K}{N}} \cdot \left(\frac{\sigma_z}{\sigma_v}\right) \tag{13b}$$

Because there are more assets than feature-specific shocks, the market maker can just run the standard OLS regression, $\phi_{\text{OLS}}(\mathbf{d},\mathbf{X})$:

$$1/\theta \cdot d_n = \mathbf{x}_n \widehat{\boldsymbol{\alpha}}_{\text{OLS}} + \epsilon_n \tag{14}$$

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

$$\text{E}\left[\frac{1}{N} \cdot \|\mathbf{v} - \mathbf{X}\widehat{\boldsymbol{\alpha}}_{\text{OLS}}\|_2^2\right] = \frac{K}{N} \cdot \frac{\sigma_z^2}{\theta^2} \tag{15}$$

Using his priors on the distribution of each asset's value, $v_n \overset{\text{iid}}{\sim} \text{N}(0,\sigma_v^2)$, 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 a really strong fundamental value realization noise trading is more erratic. 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. I now show just how hard it is to recover this information in a large market. Specifically, I show that, if the traders haven't seen at least $N^\star(Q,K)$ observations, then they will always suffer from feature-selection risk, they will always make some errors in picking which features to analyze.

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's identified which feature-specific shocks have taken place? He doesn't 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:

$$\lim_{N\to\infty} Q_N,K_N = \infty \qquad\qquad N \geq K_N \qquad\qquad \lim_{N\to\infty} {}^{K_N}/_{Q_N} = 0 \tag{16}$$

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%.[8] 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.

Formally, I am interested in the quantity:

$$\text{FSE}[\phi] = \text{E}\left[\, \left\| \text{S}[\widehat{\boldsymbol{\alpha}}] - \text{S}[\boldsymbol{\alpha}] \right\|_\infty \right] \tag{17}$$

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

$$\text{S}[\widehat{\alpha}_q] = \begin{cases} 1 & \text{if } \widehat{\alpha}_q \neq 0 \\ 0 & \text{if } \widehat{\alpha}_q = 0 \end{cases} \tag{18}$$

The ℓ_∞-norm gives a 1 if there is any difference in the support of the vectors and a 0 otherwise. In words, $\text{FSE}[\phi]$ 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, $\mathbf{X}$. Let $\boldsymbol{\Phi}$ denote the set of all possible inference strategies the market maker might use. If there exists some inference strategy $\phi \in \boldsymbol{\Phi}$ with $\text{FSE}[\phi] = 0$, then 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 given 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 $N <$

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

$N^\star(Q,K)$, 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 some constant $C > 0$ such that:*

$$N < C \cdot K_N \cdot \log(Q_N/K_N) \tag{19}$$

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

$$\min_{\phi \in \mathbf{\Phi}} \text{FSE}[\phi] > c \tag{20}$$

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

Importantly, Proposition 3.1 doesn't 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 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 wouldn't 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 derived from 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:

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

different feature combinations. Thus, $N^\star(7,7) = 7$ gives 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 isn't 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, practical application 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 estimates in the large T limit. The second step then uses the realized point estimates to compute the coefficient 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 motivating example from Section 1. Suppose that $Q = 400$, and I told you exactly which $K = 5$ of the characteristics were mispriced. For this sub-problem you could easily estimate the values of each of the coefficients using a standard regression procedure—that is, a convex approach. 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; however, 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." Current research in compressed sensing focuses on how close convex optimization programs can come to achieving this oracle bound.

4. Feature-Selection Risk

Let's now introduce feature-selection risk into the baseline asset-pricing model to see how it warps traders' perception of asset values, distorts prices, and delays arbitrage. The basic equilibrium concept will remain completely standard. The key question is 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.**

If the market maker does not have access to an oracle, then he must both identify and interpret feature-specific shocks. Since there are many more potentially relevant features than there are assets, $Q \gg N$, he must use a sparse inference strategy. No matter what sparse inference strategy he chooses, he will be subject to the feature-selection bound of Section 3. However, in order to compute equilibrium asset prices, I need to compute the market maker's posterior beliefs and this involves picking an inference strategy for him to follow.

I study a market maker who uses the least absolute shrinkage and selection operator

(LASSO) outlined in Tibshirani (1996):

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\widetilde{\boldsymbol{\alpha}}\in\mathbf{R}^Q} \left\{ \|\mathbf{X}\widetilde{\boldsymbol{\alpha}} - (1/\theta)\cdot\mathbf{d}\|_2^2 + \gamma\cdot\|\widetilde{\boldsymbol{\alpha}}\|_{\mathbf{1}} \right\} \tag{22}$$

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 $\gamma\cdot\|\widetilde{\boldsymbol{\alpha}}\|_{\mathbf{1}}$ term, then the inference strategy would be equivalent to ordinary least squares which isn't 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. 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 (2011). 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 doesn't 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 so. For example, think about the ℓ_0 problem:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\widetilde{\boldsymbol{\alpha}}\in\mathbf{R}^Q} \left\{ \|\mathbf{X}\widetilde{\boldsymbol{\alpha}} - (1/\theta)\cdot\mathbf{d}\|_2^2 + \gamma\cdot\|\widetilde{\boldsymbol{\alpha}}\|_{\mathbf{0}} \right\} \tag{23}$$

However, any 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.** I now consider the more general setting when the market maker doesn't have access to an oracle and must solve a sparse, high-dimensional, inference problem on his own. The feature-selection bound implies that the market maker now has to bear some feature-selection risk when $N < N^\star(Q,K)$. I show that informed traders in this new model earn higher profits since they can hide behind both noise trader demand shocks and feature-selection error.

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, $\mathbf{v}$, with a signal error that satisfies the inequality below:

$$\frac{1}{N}\cdot\|\mathbf{X}\widehat{\boldsymbol{\alpha}}_{\text{LASSO}} - \mathbf{v}\|_2^2 \leq \widetilde{C}^2\cdot\log(Q)\times\frac{K}{N}\cdot\frac{\sigma_z^2}{\theta^2} \tag{24}$$

with probability approaching unity as $N\to\infty$ for $\widetilde{C} = 2\cdot\sqrt{2}\cdot(1+\sqrt{2})$. Where does this $\widetilde{C}^2\cdot\log(Q)$ 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's these feature-selection errors that increase the variance of his posterior beliefs by a factor $\widetilde{C}^2 \cdot \log(Q)$ 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:

$$y_n = \theta \cdot v_n \tag{25}$$

and a linear pricing rule for the single market maker common to all N assets:

$$p_n = \lambda \cdot d_n \tag{26}$$

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:*

$$\lambda = \frac{1}{2 \cdot \theta} \tag{27a}$$

$$\theta = C \cdot \sqrt{\log(Q)} \times \sqrt{\frac{K}{N}} \cdot \left(\frac{\sigma_z}{\sigma_v}\right) \tag{27b}$$

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

Increasing the number of payout-relevant features that a market maker has to sort through, Q, delays arbitrage. First, it 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 doesn't 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 each demand gives him information about that particular asset's fundamental value, but he can't extrapolate this information from one asset to the next. Second, it makes the market maker less certain about his inferences. i.e., it imposes a penalty on precision of the market maker's posterior beliefs of $C^2 \cdot \log(Q)$ per unit of fundamental volatility for market breadth. In short, 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—that is, by

repeatedly estimating the LASSO on sample datasets. For example, when the magnitude of the K feature-specific shocks is drawn $\alpha_q \sim^{\text{iid}} \pm\text{Unif}[1,2] \cdot (\sigma_z/\theta)$, 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. Model Predictions

Let's now analyze a pair of novel empirical predictions produced by this model with feature selection risk.

5.1. **Substituting Risks.** The model outlined above predicts that it should be more profitable for an informed trader to learn a firm-specific piece of news in a markets with a larger number of payout-relevant features—that is, with a larger value of Q. After all, adding more payout-relevant features lowers the equilibrium price impact coefficient. Forcing market makers to sort through a larger number of potentially relevant features makes them less responsive to informed trader demand. To make this point precise, I study the unconditional expectation of an asset-specific value investor when the market maker has to use the LASSO to both identify and interpret feature-specific shocks:

$$\Pi(Q,\sigma_z) = \text{E}\left(\max_{y_n} \text{E}\left[\,(v_n - p_n)\cdot y_n \,|\, v_n\,\right]\right) \tag{28}$$

The proposition below shows that this quantity is increasing in the number of payout-relevant features. Put differently, a bigger haystack means more profit for the informed traders.

Proposition 5.1 (Informed Trader Profit)**.** *If the market maker does uses the LASSO to identify and interpret feature-specific shocks and $N > N^\star(Q,K)$, then the N informed traders have expected profits:*

$$\Pi(Q,\sigma_z) = {}^{C}/_{2} \cdot \sqrt{{}^{K}/_{N} \cdot \log(Q)} \times \sigma_v \cdot \sigma_z \tag{29}$$

for some positive numerical constant $0 < C < \widetilde{C}$ defined in Proposition 4.2.

What's interesting about the functional form of the informed traders' expected profits given in Proposition 5.1 is that the number of features, Q, and the volatility of noise trader demand, σ_z, enter multiplicatively. Feature-selection risk and noise trader demand risks are substitutes in a Kyle (1985)-type model. Adding more payout-relevant features for the market maker to sort through makes him less responsive to aggregate demand shocks in exactly the same way that increasing the noise trader demand volatility does.

A natural follow up question is: What is the exchange rate between feature-selection risk and noise trader demand risk? Suppose that you decreased noise trader demand volatility

by a fraction Δ_{σ_z}:

$$Q \mapsto Q' = Q \cdot (1 + \Delta_Q) \tag{30a}$$

$$\sigma_z \mapsto \sigma_z' = \sigma_z \cdot (1 + \Delta_{\sigma_z}) \tag{30b}$$

At what rate would you have to add features to the market, Δ_Q, to leave the informed traders with exactly the same profit? It turns out that for small values of Δ_{σ_z} it is possible to answer this questions. I do this by expanding the expression for the informed traders' expected profit around any baseline level of (Q,σ_z) and solving for Δ_Q as a function of Δ_{σ_z} so that the first order terms cancel out:

$$0 = \frac{\partial}{\partial Q'} \Pi(Q',\sigma_z)|_{Q'=Q} \cdot \Delta_Q + \frac{\partial}{\partial \sigma_z'} \Pi(Q,\sigma_z')|_{\sigma_z'=\sigma_z} \cdot \Delta_{\sigma_z} \tag{31}$$

The corollary below characterizes exactly this relationship.

Corollary 6.1 (Substituting Risks)**.** *Suppose you decreased the noise trader demand volatility by a fraction $\Delta_{\sigma_z} > 0$, then increasing the number of asset features by a fraction:*

$$\Delta_Q = 2 \cdot \log(Q) \cdot \left(\frac{Q}{\sigma_z}\right) \times \Delta_{\sigma_z} \tag{32}$$

would leave informed trader expected profits and the price impact coefficient, λ, unchanged.

5.2. **Seemingly Redundant Assets.** The model outlined above also predicts that the market maker can identify feature-specific shocks using fewer assets if he studies a collection of assets with a more random assortment of feature exposures. The textbook approach to learning about risks involves studying the prices of simple assets: 1 Arrow security for each shock. By contrast, compressed sensing theory asserts that an astute trader can identify feature-specific shocks from the prices of far fewer assets if *a*) the shocks are sparse and *b*) the chosen assets have extremely heterogeneous exposures to a large number of features.

One way to learn about feature-specific shocks is to look at the price and demand of Arrow securities. For example, if there are Q payout relevant features:

$$\begin{bmatrix} d_1^{(A)} \\ d_2^{(A)} \\ d_3^{(A)} \\ \vdots \\ d_Q^{(A)} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}}_{\mathbf{x}^{(A)}} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_q \end{bmatrix} + \frac{1}{\theta} \cdot \begin{bmatrix} z_1^{(A)} \\ z_2^{(A)} \\ z_3^{(A)} \\ \vdots \\ z_Q^{(A)} \end{bmatrix} \tag{33}$$

This market setup is incredibly simple. The aggregate demand for the first Arrow security, $d_1^{(A)}$, tells the market maker if there has been a feature-specific shock to the first feature; the aggregate demand for the second Arrow security, $d_2^{(A)}$, tells the market maker if there has been a feature-specific shock to the second feature; the aggregate demand for the third

Arrow security, $d_3^{(A)}$, tells the market maker if there has been a feature-specific shock to the third feature; and so on...

Arrow securities are simple, but they are also wasteful. They don't exploit the fact that the market maker knows $\boldsymbol{\alpha}$ is spiky and concentrated in only a few of its coordinates. Arrow securities are informative because they form an orthonormal basis. As a result, no 2 sets of feature-specific shocks can manifest themselves to the market maker in aggregate demand in exactly the same way. Yet, the market maker doesn't care about all possible collections of feature-specific shocks. He just cares about K-sparse shocks. Asset complexity gives a way for the market maker to exploit his knowledge of the sparsity of the feature-specific shocks.

For example, consider a collection of N derivative assets constructed by financial engineers out of the Q Arrow securities. These derivative assets will have an $(N \times Q)$-dimensional exposure matrix $\mathbf{X}$:

$$\underset{N\times Q}{\mathbf{X}} = \overset{N\times Q}{\mathbf{D}}\,\underset{Q\times Q}{\mathbf{X}^{(A)}} \tag{34}$$

Obviously, the N derivative assets can't have completely independent exposure to each of the Q payout-relevant features since $N \ll Q$. Some of the derivatives will have to have similar exposures to, say, crowded trade risk and S&P 500 inclusion risk. However, the market maker doesn't need the derivatives to be a completely linearly independent set of risk exposures. He just needs them to be sufficiently different.

Specifically, suppose that any $(2 \cdot K)$ columns of the $(N \times Q)$-dimensional derivative feature-exposure matrix $\mathbf{X}$ are linearly independent. Then, any K-sparse signal $\boldsymbol{\alpha} \in \mathbf{R}^Q$ can be reconstructed uniquely from $\mathbf{X}\boldsymbol{\alpha}$. If not, then there would have to be a pair of K-sparse signals $\boldsymbol{\alpha}$,$\boldsymbol{\alpha}' \in \mathbf{R}^Q$ with $\mathbf{X}\boldsymbol{\alpha} = \mathbf{X}\boldsymbol{\alpha}'$; however, this would imply that $\mathbf{X}(\boldsymbol{\alpha} - \boldsymbol{\alpha}') = 0$ which is a contradiction. $\boldsymbol{\alpha} - \boldsymbol{\alpha}'$ is at most $(2 \cdot K)$-sparse, and there can't be a linear dependence between $(2 \cdot K)$ columns of $\mathbf{X}$ by assumption. Thus, the market maker is happy to tolerate a little bit of redundancy. So long as traders can replicate the market's exposure to any $(2 \cdot K)$ features with fewer than N assets, aggregate demand shocks to the N assets will reveal which K feature-specific shocks have occurred.

It is possible to generalize this result to random matrices. For an $(N \times Q)$-dimensional matrix $\mathbf{X}$, the K-restricted isometry constant δ_K is the smallest number such that:

$$\max_{|\mathcal{J}|\leq K} \|{}^1\!/_N \cdot \mathbf{X}_{[\mathcal{J}]}^\top \mathbf{X}_{[\mathcal{J}]} - \mathbf{I}\|_2 \leq \delta_K \tag{35}$$

For matrices with small restricted isometry constants, every subset of K or fewer columns is approximately an orthonormal system. Clearly, choosing $\mathbf{X}^{(A)} = \mathbf{I}$ via Arrow securities means that $\delta_K = 0$; however, Candes and Tao (2005) show that matrices with Gaussian entries, $x_{n,q} \overset{\text{iid}}{\sim} \text{N}(0,1)$, have small restricted isometry constants and allow for K-sparse recovery with very high probability whenever the number of measurements N is on the order

of $N^{\star}(Q,K) = K \cdot \log(Q/K)$. Proposition 5.2 characterizes the savings in required observations from examining complex derivative assets rather than Arrow securities.

Proposition 5.2 (Seemingly Redundant Assets)**.** *If $N \geq N^{\star}(Q,K)$, then a market maker using the LASSO with $\gamma = 2 \cdot (\sigma_z/\theta) \cdot \sqrt{2 \cdot \log(Q)}$ to study the aggregate demand for complex derivatives whose feature exposures are drawn $x_{n,q} \overset{\text{iid}}{\sim} \text{N}(0,1)$ can identify a K-sparse set of feature-specific shocks with probability greater than $1 - C_1 \cdot e^{-C_2 \cdot K}$ using:*

$$\Theta[K/Q \cdot \log(Q/K)] \tag{36}$$

times fewer assets than a market maker studying the aggregate demand for Arrow securities on each of the Q features where $C_1, C_2 > 0$ are numerical constants.

There is an interesting analogy to randomized control trials here. i.e., randomizing which assets get sold makes price changes and demand schedules more informative about feature-specific shocks in the same way that randomizing which subjects get treated in a medical study makes the experimental results more informative about the effectiveness of a drug. Why does randomization help? Suppose all of the people who got the real drug recovered and all of the people who got the placebo didn't. Randomly assigning patients to the treatment and control groups makes it exceptionally unlikely that the patients who took the real drug will happen to have some other trait (e.g., a genetic variation) that actually explains their recovery. Randomizing feature exposures decreases the probability that 2 different K-sparse vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ are observationally equivalent when looking only at public market data.

6. 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. e.g., papers like Sims (2006) and Hong, Stein, and Yu (2007) suggest that cognitive costs force traders to use overly simplified mental models, and Gabaix (2011) 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 formulation relies on the fact that asset values are governed at least in part by a constantly changing cast of feature-specific shocks. Chinco (2014) 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). e.g., 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. e.g., 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 F.A. Hayek's notion of local knowledge. Indeed, Hayek (1945) gives trader who benefits from specialized experience with particular assets as a canonical example of a situation requiring 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 (e.g., see Hong and Stein (2007)) and noise trading (e.g., see Black (1986)). i.e., 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. I investigate these ideas further in Appendix E.

7. 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 act like a limit to arbitrage 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. *The Journal of Finance 61*(1), 259–299.

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

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

Black, F. (1986). Noise. *The 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. *The 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? *The Review of Financial Studies 25*(9), 2745–2787.

Chinco, A. (2014). No coincidence, no story. *Working Paper*.

Cohen, L. and A. Frazzini (2008). Economic links and predictable returns. *The 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.

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

Donoho, D. and J. Jin (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *The Annals of Statistics 34*(6), 1593–3050.

Gabaix, X. (2011). A sparsity-based model of bounded rationality. *NBER Working Paper* (16911).

Grossman, S. and J. Stiglitz (1980). On the impossibility of informationally efficient markets. *American Economic Review 70*(3), 393–408.

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

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2014). The common factor in idiosyncratic volatility. *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. *The 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. *The Journal of Finance 56*(1), 387–396.

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 data-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 financial asset pricing models. *The Review of Financial Studies 3*(3), 431–467.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics 37*(1), 246–270.

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

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

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

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

Sullivan, R., A. Timmermann, and H. White (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The 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.

Veldkamp, L. (2006). Information markets and the comovement of asset prices. *Review of Economic Studies 73*(3), 823–845.

Veldkamp, L. (2011). *Information Choice in Macroeconomics and Finance* (1 ed.). Princeton University Press.

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.

Weisberg, S. (2005). *Applied Linear Regression* (2 ed.). John Wiley & Sons.

Appendix 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:

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

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 $\mathbf{X}_{[\mathcal{K}]}$ denotes the measurement matrix $\mathbf{X}$ 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 $\boldsymbol{\alpha}$:

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

Thus, the cross-section of aggregate demand gives the market maker a signal about each asset's fundamental value:

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

which has signal error:

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

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:

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

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,* $\mathrm{Cor}[x,y]\neq 0$*, and let $f(y)$ be the predicted value of x for some deterministic function $f(\cdot)$. Then we have that:*

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

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:

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

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 7 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. i.e., 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:

$$p_s = \mathrm{N}(\alpha_{\min} \cdot \mathbf{X}[\mathcal{K}_s]\mathbf{1},\mathbf{I}) \qquad \text{for } s = 1,2,\ldots,S$$

where $\mathbf{X}[\mathcal{K}_s]$ denotes the observed measurement matrix restricted to the columns $\mathcal{K}_s$, $\mathbf{1}$ denotes a $(K \times 1)$-dimensional vector of 1s, and $\mathbf{I}$ 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:

$$\mathrm{FSE}[\phi] = \Pr[\mathcal{K} = \phi(\mathbf{d},\mathbf{X})] \geq 1 - \frac{\mathrm{M}[p,\mathbf{d}|\mathbf{X}]}{\log_2(S)} - \mathrm{o}(1)$$

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 $\mathrm{M}[p,\mathbf{d}|\mathbf{X}]$ which can be upper bounded as follows:

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

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

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

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

$$\mathrm{FSE}[\phi] = \Pr[\mathcal{K} = \phi(\mathbf{d},\mathbf{X})] \geq 1 - \left(\frac{\frac{1}{2 \cdot S^2} \cdot \sum_{s,s'=1}^{S} \|\alpha_{\min} \cdot (\mathbf{X}[\mathcal{K}_s] - \mathbf{X}[\mathcal{K}_{s'}])\mathbf{1}\|_2^2}{\log_2(S)} \right) - \mathrm{o}(1)$$

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

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

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

variable:

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

Because each asset has feature exposures, $x_{n,q} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, $h_{s,s'}$ follows a χ^2_N distribution:

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

where $|\mathcal{K}_s \cap \mathcal{K}_{s'}|$ denotes the size of the set difference between the subsets $\mathcal{K}_s$ and $\mathcal{K}_{s'}$. e.g., 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:

$$\Pr\left[\frac{1}{S^2} \cdot \sum_{s \neq s'} h_{s,s'} \geq 4 \cdot \alpha_{\min}^2 \cdot K \cdot N\right] \leq 1/2$$

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

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

(6) *Formulate key inequality.* Thus, as long as:

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

the error rate will remain bounded away from 0 implying that:

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

is necessary for $\mathrm{FSE}[\phi] \to 0$. The multiplier $(4 \cdot \alpha_{\min}^2 \cdot K)^{-1}$ 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.

(7) *Make cosmetic touch-up.* 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, $\widehat{\boldsymbol{\alpha}}_{\text{LASSO}}$, from the program in equation (22) using the tuning parameter $\gamma = 2 \cdot (\sigma_z/\theta) \cdot \sqrt{2 \cdot \log(Q)}$ obeys:*

$$\Pr\left[\frac{1}{N} \cdot \|\mathbf{X}\boldsymbol{\alpha} - \mathbf{X}\widehat{\boldsymbol{\alpha}}\|_2^2 \leq \widetilde{C}^2 \times \left(\frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta^2}\right)\right] \geq 1 - \frac{6}{Q^{2 \cdot \log 2}} - \frac{1}{Q \cdot \sqrt{2 \cdot \pi \cdot \log(Q)}}$$

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

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:

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

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

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

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 $\mathbf{v}$ and variance given in Lemma A.3.

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

$$\text{Var}[v_n|\mathbf{d}] = \left(\frac{C^2 \cdot \frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}}{\sigma_v^2 + C^2 \cdot \frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}} \right) \times \sigma_v^2 \qquad \text{E}[v_n|\mathbf{d}] = \underbrace{\frac{1}{\theta} \cdot \left(\frac{\sigma_v^2}{\sigma_v^2 + C^2 \cdot \frac{K \cdot \log(Q)}{N} \cdot \frac{\sigma_z^2}{\theta(\lambda)^2}} \right)}_{\lambda} \cdot \mathbf{d}$$

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

Proof (Proposition 5.1). Plugging the price impact and demand coefficients from Proposition 4.2 into the informed trader's optimization program in Equation 7 gives:

$$\begin{aligned} \Pi(Q,\sigma_z) &= \text{E}\left(\max_{y_n} \text{E}\left[\left(v_n - \lambda \cdot \{y_n + z_n\}\right) \cdot y_n \,|\, v_n \right] \right) \\ &= \text{E}\left[\left(v_n - \lambda \cdot \{\theta \cdot v_n + z_n\}\right) \cdot \theta \cdot v_n \right] \\ &= \theta \cdot \frac{\sigma_v^2}{2} \end{aligned}$$

Setting $\theta = C \cdot \sqrt{\log(Q)} \times \sqrt{{}^K/_N} \cdot ({}^{\sigma_z}/_{\sigma_v})$ and simplifying gives the desired result. □

Proof (Corollary 6.1). The functional form of the informed trader's expected profits comes from Proposition 5.1. Its partial derivative with respect to the number of features is given by:

$$\frac{\partial}{\partial Q'} \Pi(Q',\sigma_z)|_{Q'=Q} = \frac{C}{2} \times \left(\frac{1}{2} \cdot \frac{1}{\sqrt{\frac{K}{N} \cdot \log(Q)}} \right) \times \left(\frac{K}{N \cdot Q} \right) \times \sigma_v \cdot \sigma_z$$

Its partial derivative with respect to the amount of noise trader demand volatility is given by:

$$\frac{\partial}{\partial \sigma_z'} \Pi(Q,\sigma_z')|_{\sigma_z'=\sigma_z} = \frac{C}{2} \cdot \sqrt{\frac{K}{N} \cdot \log(Q)} \cdot \sigma_v$$

In order for a tiny increase in the number of features, Δ_Q, to offset a tiny decrease in the amount of noise trader demand volatility, Δ_{σ_z}, the following condition has to hold:

$$\Delta_Q \times \frac{C}{2} \times \left(\frac{1}{2} \cdot \frac{1}{\sqrt{\frac{K}{N} \cdot \log(Q)}} \right) \times \left(\frac{K}{N \cdot Q} \right) \times \sigma_v \cdot \sigma_z = -\Delta_{\sigma_z} \times \frac{C}{2} \cdot \sqrt{\frac{K}{N} \cdot \log(Q)} \cdot \sigma_v$$

Simplifying then yields the desired result. □

Lemma A.4 (Bound on LASSO Recovery Error, Wainwright (2009b)). *If $N \geq N^\star(Q,K)$, then the LASSO estimate, $\widehat{\boldsymbol{\alpha}}_{\text{LASSO}}$, from the program in equation (22) using the tuning*

parameter $\gamma = 2 \cdot (\sigma_z/\theta) \cdot \sqrt{2 \cdot \log(Q)}$ *identifies the correct subset of feature-specific shocks with probability greater than:*

$$1 - C_1 \cdot \exp\{-C_2 \cdot K\}$$

for numerical constants C_1,$C_2 > 0$.

Proof (Proposition 5.2). First, consider the market maker studying Arrow securities that each have exposure to exactly 1 feature. The probability that any particular Arrow security will realize a shock is K/Q. The expected number of securities he needs to investigate before he sees all K feature-specific shocks is then given by the mean of negative binomial distribution with K failures:

$$\frac{(1 - K/Q) \cdot K}{K/Q} \asymp Q$$

Second, consider the market maker studying complex derivatives. Lemma A.4 says that he can identify the correct features with exceedingly high probability using only $K \cdot \log(Q/K)$. The quotient gives the desired result. □

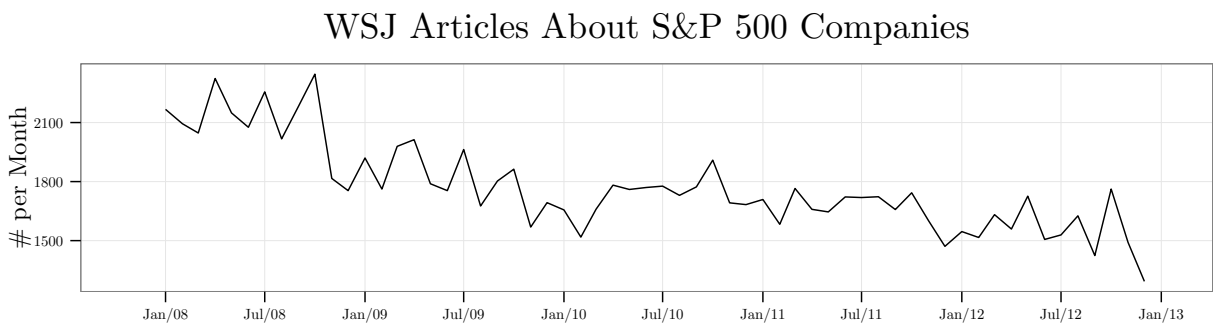

Figure 2. *Number of Wall Street Journal articles about S&P 500 companies per month from January 2008 to December 2013. Reads: "There were roughly* 1800 *articles written about S&P 500 companies in the Wall Street Journal in July 2010."*

Appendix B. Counting Asset Features

A common question people have is: Is it possible to count the number of asset features, Q, in a market? Yes. I examine Wall Street Journal article keywords. The universe of keywords ever used is an estimate of the number of features. Even after controlling for the number of news articles, the number of asset features can vary by 2 orders of magnitude for S&P 500 stocks. The data are hand-collected from the ProQuest newspaper archive.[9] The resulting data set contains 106k articles over 5 years concerning 542 companies. Many articles reference multiple S&P 500 companies. Figure 2 plots the total number of articles in the database per month. There is a steady downward trend. The first part of the sample was the height of the financial crisis, so as markets have calmed down journalists have devoted fewer articles to corporate news relative to other things such as politics and sports.

[9]See the online supporting materials at http://www.alexchinco.com/wsj-article-subject-tags/ for more details.

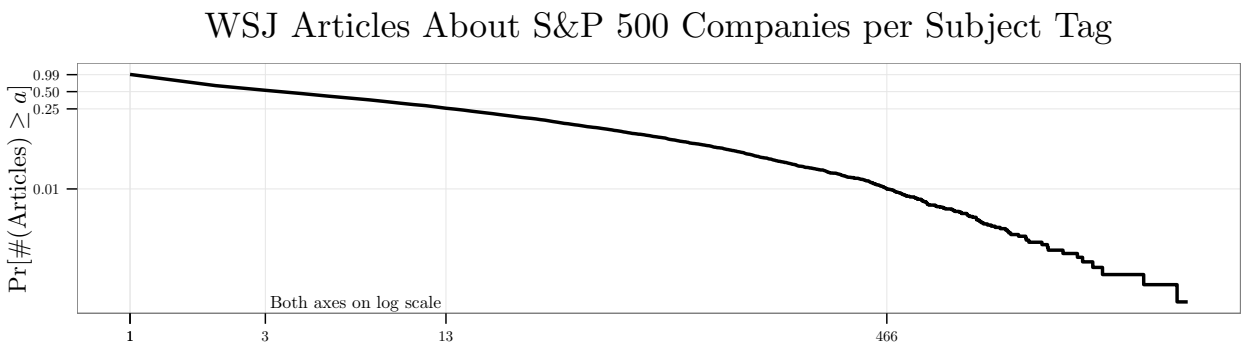

Figure 3. *Number of Wall Street Journal articles per subject tag in articles about S&P 500 companies from January 2008 to December 2013. x-axis: Number of Wall Street Journal articles. y-axis: Fraction of all subject tags used in at least that many articles. Both axes are on a logarithmic scale. The break points on the y-axis define the* 1%*,* 25%*,* 50%*, and* 99% *quantiles. Reads: "While* 50% *of all subject tags are used in* 3 *or fewer articles, the most common* 1% *of the subject tags get used in* 466 *or more articles."*

Consistent with idea that companies get hit with new and different kinds of feature-specific shocks, Figure 3 shows that the vast majority of subject tags during the sample are only used in a couple of articles. While 50% of all subject tags are used in 3 or fewer articles, the most common 1% of the subject tags get used in 466 or more articles. Traders have to figure out which aspect of the company matters. This is clearly not an easy problem to solve. Lot's of ideas are thrown around. Many of them must be either short lived or wrong. Roughly 1 out of every 4 topics worth discussing is only worth discussing once.

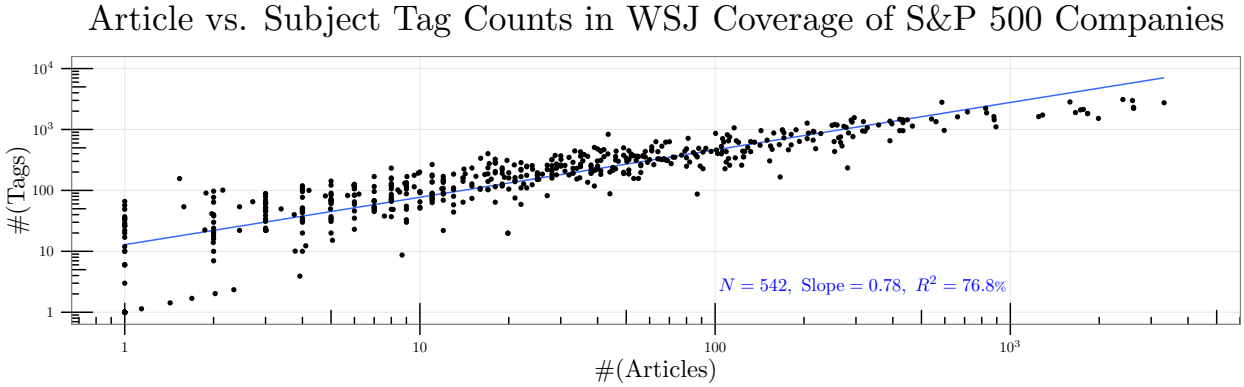

Figure 4. *Number of Wall Street Journal articles about each S&P 500 company (x-axis) vs. number of unique subject tags used to describe each S&P 500 company (y-axis) over the period from January 2008 to December 2013. Both axes are on a logarithmic scale. Reads: "S&P 500 companies with between* 100 *and* 200 *articles in the Wall Street Journal typically have anywhere between* 200 *and* 1000 *distinct subject tags."*

In addition, I find that there is substantial heterogeneity in how many different topics people write about when discussing a company even after controlling for the number of total articles as shown in Figure 4. e.g., there were 87 articles in the Wall Street Journal referencing Garmin (GRMN) and 81 articles referencing Sprint (S); however, while there were only 87 different subject tags used in the articles about Garmin, there were 716 different subject tags used in the articles about Sprint! This finding is consistent with the idea that some firms

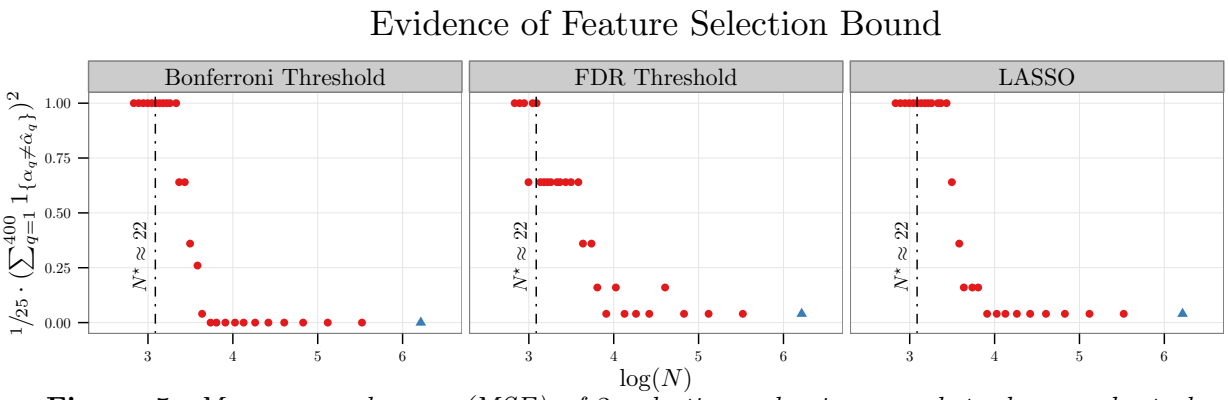

Figure 5. *Mean squared error (MSE) of* 3 *selection rules in a market where each stock has* $Q = 400$ *features and the market realizes only* $K = 5$ *feature-specific shocks as the number of observations increases from* $N = 15$ *to* $N = 400$. *Left: Traders run univariate regressions and keep variables with t-stats exceeding* $\sqrt{2 \cdot \log Q} \approx 3.46$. *Middle: Traders use same regression procedure, but keep variables with p-values less than* $0.25 \cdot (\widehat{K}/Q)$ *where* $\widehat{K}$ *is the number of data-implied parameters in the model. Right: Traders select features using LASSO. Reads: "All* 3 *procedures display a sudden drop in MSE at* $N^\star(400{,}5) \approx 22$."

face a much wider array of shocks than others. i.e., the width of the market matters.

Appendix C. Standard Inference Problem

Traders in most information-based asset-pricing models solve a Gaussian inference problem as in DeGroot (1969). e.g., market makers see N signals, d_n, which are informative about a fixed mean, $\bar{\alpha}$, that is contaminated with some noise, $\epsilon_n \overset{\text{iid}}{\sim} \text{N}(0{,}\sigma_\epsilon^2)$. We might think about these signals as excess demand telling us about market sentiment à la Baker and Wurgler (2006):

$$d_n = \tilde{d}_n - \text{E}[\tilde{d}_n|\mathbf{f}] = \bar{\alpha} + \epsilon_n \tag{37}$$

where d_n denotes asset n's excess demand and $\mathbf{f}$ denotes a vector of factors. This framework has been extremely popular and productive because it leads to simple, intuitive, closed-form solutions. e.g., if traders have prior beliefs, $\bar{\alpha} \overset{\text{iid}}{\sim} \text{N}(0{,}\sigma_{\bar{\alpha}}^2)$, then their beliefs about $\bar{\alpha}$ after seeing N signals are given by:

$$\text{Var}[\bar{\alpha}|\mathbf{d}] = \sigma_{\bar{\alpha}}^2 \cdot \left(\frac{\sigma_\epsilon^2}{N \cdot \sigma_{\bar{\alpha}}^2 + \sigma_\epsilon^2}\right) \qquad \text{and} \qquad \text{E}[\bar{\alpha}|\mathbf{d}] = \left(\frac{\sigma_{\bar{\alpha}}^2}{N \cdot \sigma_{\bar{\alpha}}^2 + \sigma_\epsilon^2}\right) \cdot \sum_{n=1}^{N} d_n \tag{38}$$

See Veldkamp (2011) for an excellent overview of this literature.

Appendix D. Numerical Example

Suppose that stocks have $Q = 400 \gg 7$ features and the market realizes $K = 5 > 1$ feature-specific shocks so that aggregate demand is given by:

$$d_n = \tilde{d}_n - \text{E}[\tilde{d}_n|\mathbf{f}] = \sum_{q=1}^{400} \alpha_q \cdot x_{n,q} + \epsilon_n \qquad \text{and} \qquad 5 = \|\boldsymbol{\alpha}\|_{\ell_0} = \sum_{q=1}^{400} 1_{\{\alpha_q \neq 0\}} \tag{39}$$

Here, $x_{n,q} \overset{\text{iid}}{\sim} \text{N}(0{,}1)$ denotes stock n's exposure to the qth feature, $\epsilon_n \overset{\text{iid}}{\sim} \text{N}(0{,}\sigma_\epsilon^2)$ denotes idiosyncratic noise for stock n, and $\alpha_q = 1/\sqrt{K}$ for all $q \in \{q' \in \mathcal{Q} : \alpha_{q'} \neq 0\}$. Notice that in

this extension, I am no longer hand-picking each stocks feature exposures.

There are a number of statistical techniques to identify which 5 of the 400 features have realized a shock. First, you might try forward stepwise regression as in Weisberg (2005),

$$d_n = \widehat{\alpha}_q \cdot x_{n,q} + \varsigma_n \qquad \text{for all } q = 1,2,\ldots,Q \tag{40}$$

keeping only the variables whose t-statistics exceed the Bonferroni threshold of $\sqrt{2 \cdot \log Q} \approx 3.46$. The left panel of Figure 5 shows the mean squared error from this approach. As you would expect, if there are more observations for you to analyze—that is, moving left to right, then you are better able to identify the 5 shocked features. However, the change doesn't happen gradually. Your error rate in interpreting aggregate demand schedules suddenly plummets once you've seen $N^\star(400,5) \approx 22$ observations. This is the feature-selection bound.

Here is the interesting part. This critical number is independent of the statistical procedure you use. For example, the middle panel shows the results if you were to use the same stepwise-regression procedure but keep only the variables whose p-values were less than $0.25 \cdot (\widehat{K}/Q)$, with $\widehat{K}$ denoting the total number of data-implied parameters in the model. This cutoff is known as the false-discovery-rate (FDR) threshold and comes from Donoho and Jin (2006). Alternatively, the right panel shows the results if you were to use the least absolute-shrinkage and selection operator (LASSO) as in Tibshirani (1996). Each panel displays a sudden drop in the error rate just after the feature-selection bound has been reached. The bound is a generic property of the high-dimensional inference problem.

Appendix E. Local Knowledge

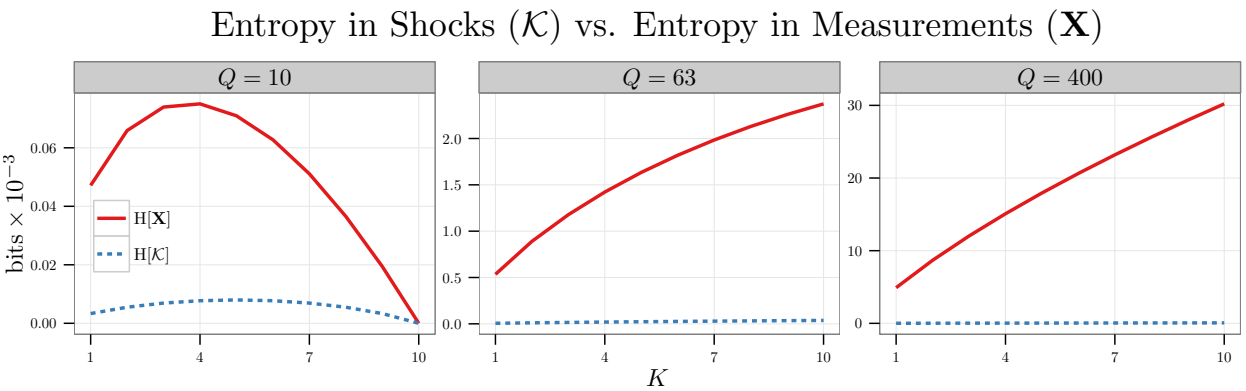

Figure 6. *Entropy needed to transmit both the choice of K feature-specific shocks,* $\mathrm{H}[\mathcal{K}]$ *(blue, dashed), and the feature-exposure matrix for the $N^\star$ observations needed to identify them,* $\mathrm{H}[\mathbf{X}]$ *(red, solid), as the number of shocks grows from $K = 1$ to $K = 10$ for $Q \in \{10,63,400\}$. Reads: "$\mathrm{H}[\mathbf{X}] = 18 \times 10^3$ bits and $\mathrm{H}[\mathcal{K}] = 36$ bits when $Q = 400$ and $K = 5$ corresponding to a vertical line through the right panel at $K = 5$. Thus, it takes 500 times as much information to measure all of the feature exposures for the $N^\star \approx 22$ assets needed to identify $\mathcal{K}$ as it does to record the actual configuration entropy of $\mathcal{K}$."*

I conclude this paper by examining the role of local knowledge in this analysis. The goal in this subsection is to shed light on how local knowledge differs from the usual notions of cognitive costs in the economics and finance literature. e.g., in existing information-based asset-pricing models, the cost of a signal typically scales with how much smarter it makes you as measured by an increase in the precision of your posterior beliefs or a reduction in their

entropy. e.g., see Veldkamp (2006) for a representative example. By contrast, in the current paper the cost of acquiring knowledge about which K features have realized a shock scales with the number of measurements necessary to uncover this information. What's more, the entropy bound up in these measurements typically exceeds the actual entropy of the signal by an order of magnitude or more. I call this gap the amount of local knowledge in the market.

To make these statements more precise, I first calculate how much information it would take to convey which K feature-specific shocks have occurred in a market with Q characteristics and K shocks. Let $\binom{Q}{K} = W$ denote the number of ways to select K characteristics from among Q possibilities. The amount of information in the signal is then given by the configuration entropy of $\mathcal{K}$ in units of bits:

$$\mathrm{H}[\mathcal{K}] = -\sum_{w=1}^{W} \frac{1}{W \cdot \log(2)} \cdot \log\left(\frac{1}{W}\right) \tag{41}$$

Yet, in order for a market maker to uncover this signal by studying the cross-section of aggregate demand, he has to observe the feature exposures of $N^\star(Q,K)$ assets. This is a $(N^\star \times Q)$-dimensional matrix with elements $x_{n,q} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$. Each element in this matrix is a single measurement. Thus amount of information necessary to store all of these measurements in units of bits is given by:

$$\mathrm{H}[\mathbf{X}] = -\frac{1}{2 \cdot \log(2)} \cdot \log\left[\frac{1}{(2 \cdot \pi \cdot e)^{Q \cdot N^\star}}\right] \tag{42}$$

The proposition below characterizes how the value to market-wide arbitrageurs of immediately discovering the local knowledge scales with the signal recovery bound.

Proposition D (Local Knowledge)**.** *The entropy of the measurements needed to discover which K feature-specific shocks have occurred exceeds the configuration entropy of the shocks:*

$$\text{Local Knowledge} = \mathrm{H}[\mathbf{X}] - \mathrm{H}[\mathcal{K}] \geq 0 \tag{43}$$

Proof. $\mathrm{H}[\mathcal{K}]$ is maximized with shock probability for each of the Q features is $1/2$. The entropy of Q independent random normal variables is at least as large as that of a binomial distribution with Q draws and probability $1/2$. $N^\star(Q,K) \geq 1$. □

In order to uncover which K feature-specific shocks have occurred, a market maker has to observe the prices of a bunch of assets with randomly assigned feature exposures so that it is really unlikely that 2 different sets of K-sparse shocks can produce the same pattern in aggregate demand by pure chance. This means that the feature-exposure matrix, $\mathbf{X}$, is a big, random, unstructured matrix. As a result, it takes a lot of entropy to store it. e.g., when $Q = 400$ and $K = 5$, it takes 500 times as much information to measure all of the feature exposures for the $N^\star \approx 22$ assets needed to identify $\mathcal{K}$ as it does to record the actual configuration entropy of $\mathcal{K}$ as shown in Figure 6.