Equilibrium Data Mining and Data Abundance*

Jérôme Dugast† Thierry Foucault‡

November, 2021

Abstract

We analyze how computing power and data abundance affect speculators' search for predictors. Speculators optimally stop searching when they find a predictor with a signal-to-noise ratio greater than an endogenous threshold. Greater computing power raises this threshold by reducing search costs. In contrast, data abundance can reduce this threshold because (i) it reduces rents from informed trading, except for the best informed speculators and (ii) it increases the average number of trials required to find a predictor. We derive predictions regarding the effects of progress in information technologies on active asset managers' performance, the similarity of their holdings and the informativeness of asset prices.

Keywords: Alternative Data, Data Abundance, Data Mining, Price Informativeness, Search for Information.

*We are grateful to Simona Abis, Snehal Banerjee, Bruno Biais, Dion Bongaerts, Maxime Bonnelli, Adrian Buss, Jean-Edouard Colliard, Bernard Dumas, Maryam Farboodi, Sergei Glebkin, Denis Gromb, Johan Hombert, Pete Kyle, Sophie Moinas, Joël Peress, Francesco Sangiorgi, Daniel Schmidt, Andriy Shkilko (discussant), Alberto Teguia, Mao Ye, Josef Zechner and participants at the ILB Rising Talents in Finance and Insurance Conference, the Paris Finance December Meeting 2020, the 2020 European Finance Association Meetings, the 2020 future of Financial Information Conference, the 2021 Western Finance Association Meetings, the 2021 FIRS Conference, the Microstructure Exchange virtual seminar, Frankfurt School of Management, HEC Paris, INSEAD, Mc Gill, the University of Maryland, the University of New South Wales, and the University of Vienna for very useful comments. All errors are the authors alone.

†Université Paris-Dauphine, Université PSL, CNRS, DRM, Finance, 75016 PARIS, FRANCE. Tel: (+33) 01 44 05 40 41 ; E-mail: jerome.dugast@dauphine.psl.eu

‡HEC, Paris and CEPR. Tel: (33) 1 39 67 95 69; E-mail: foucault@hec.fr

1. Introduction

Active asset managers play a central role in securities markets. They make substantial investments to produce information about asset payoffs and, by trading on their information, they make securities prices more informative. Technological progress is changing how asset managers obtain information by enabling them to use (i) more diverse data (generated by social media, web search, online transactions, mobile phones etc.) and (ii) more powerful computer-based methods (e.g., machine learning) to mine these data.[1] As a result, the set of potential predictors for asset payoffs has considerably increased.[2] This evolution raises many interesting questions: How does it affect managers' search for predictors? How does it affect asset managers' performance? Does it make asset managers' signals and holdings more or less similar? Does it make asset prices more informative?

In this paper, we address these questions. A unique feature of our theory is to differentiate the expansion of the set of available data to find predictors (data abundance) from the reduction of information processing costs (greater computing power). These two aspects of the big data revolution are related but conceptually distinct.[3] To develop predictions about the effects of progress in information technologies, it is therefore important to develop models of information acquisition, like ours, in which data abundance and computing power are different parameters.

Our model features a continuum of risk averse speculators (which we interpret as quantitative asset managers). In the first stage (the "exploration stage"), each speculator optimally scours available data to select predictors of a risky asset payoff. In the second stage (the "trading stage"), each speculator observes the realization of her predictor and optimally trades on this information, as in other rational expectations models (the trading stage is similar to Vives (1995)). The novel implications of our model stem from the exploration stage. Here, instead of following the standard approach (e.g., Grossman and Stiglitz (1980) or Verrecchia (1982)), whereby speculators obtain a predictor of a given

[1] See Goldman Sachs Asset Management (2016): "*The role of big data in investing.*" Marenzi (2017) estimates that asset managers have spent more than four billion in alternative data in 2017 (see also "*Asset managers double spending in new data in hunt for edge*", Financial Times, May 9, 2018).

[2] For instance, Martin and Nagel (2020) note (on p.2) that: "*As technology has improved, the set of available and potentially valuation-relevant predictor variables has expanded enormously over time*" while Gu et al. (2020) write (on p.2225) that machine learning enables investors to use "*efficient algorithms for searching among a vast number of potential model specifications.*

[3] For instance, social media or geolocation data expands the set of variables that investors can consider to find predictors but does not per se lower the cost of processing these data).

precision at a fixed cost, we explicitly model the search for a predictor and we analyze how the optimal search strategy depends on (i) the cost of exploration (computing power) and (ii) the "search space" (data abundance).

We model the search for predictors as follows. Each speculator can combine variables (e.g., past returns, accounting variables and social media data) from different datasets to build predictors. A predictor is characterized by its signal-to-noise ratio ("quality"). The quality of a given predictor is a priori unknown but speculators know the distribution of quality across predictors. The lowest quality is zero (just noise) while the highest quality determines the "data frontier," denoted τ^{max}. Given this distribution, each speculator simultaneously and independently discovers predictors during the exploration phase. Discovering a predictor and its quality requires launching a round of exploration, which costs "c." A round returns a predictor with probability α and fails otherwise. After obtaining a predictor, a speculator can decide to trade on the predictor or to search for another one, which requires paying the exploration cost again. This process goes on until the speculator finds a satisficing predictor.

In practice, discovering and selecting a predictor requires (a) buying and preparing new data for analysis, (b) building a predictor with these data and assessing its quality with statistical techniques and (iii) deciding, via extensive backtesting, whether a predictor is good enough for live trading.[4] One round of exploration comprises all these tasks and we interpret the exploration cost as the total cost of executing them (which includes labor and opportunity costs). We assume that automation and greater computing power reduces this cost because they allow to complete an exploration round faster.[5]

In contrast, we assume that data abundance affects the distribution of predictors' quality in two ways. Firstly, data abundance enables speculators to discover new predictors by using combination of variables that previously were not available (e.g., data from social media). This possibility pushes further the data frontier, τ^{max}, i.e., improves the quality of the best predictors (the "hidden gold nugget" effect).[6] Secondly, data abun-

[4]See Chapters 8 and 9 in Narang (2013) for a practitioner's account of the way quant funds generate predictors.

[5]Brogaard and Zareei (2019) use a genetic algorithm approach to select technical trading rules. They note that "*the average time needed to find the optimum trading rules for a diversified portfolio of ten NYSE/AMEX volatility assets for the 40 year sample using a computer with an IntelÂ® Core(TM) CPU i7-2600 and 16 GM RAM is 459.29 days (11,022.97 hours)." For one year it takes approximately 11.48 days.*" They conclude that their analysis would not be possible without the considerable increase in computing power in the last 20 years.

[6]This effect is often discussed in the financial press (e.g., "*Hedge funds see a gold rush in data mining*",

dance creates a "needle in the haystack" problem: It results in a proliferation of datasets, among which only a fraction is useful for forecasting asset payoffs. This effect increases the likelihood that a particular dataset proves useless after being tested, i.e., reduces α. In our model, we analyze these two effects separately by varying either τ^{max} or α.

In equilibrium, each speculator optimally stops searching for a predictor after finding one whose quality exceeds an endogenous threshold, τ^*. This threshold is such that the speculator's expected utility of trading on a predictor of quality τ^* is just equal to her expected utility of searching for another predictor (her continuation value). Thus, the quality of predictors used in equilibrium ranges from τ^* (least informative) to τ^{max} (most informative). Hence, even though speculators are ex-ante homogeneous (same preferences and exploration costs), they are heterogeneous in the quality of their predictors (and therefore performance) due to serendipity in search outcomes.

Data abundance and computing power do not have the same effects on speculators' optimal search policy, τ^*. To understand why, it is useful to contrast the effect of a decrease in the cost of exploration, c, and the effect of an increase in the quality of the best predictor, τ^{max} on the value of launching a new round of exploration after finding a predictor (the continuation value), holding the search policy (τ^*) constant. An increase in τ^{max} has two countervailing effects. On the one hand, it raises the continuation value because the expected utility of trading on the best predictor becomes even larger. On the other hand, speculators who obtain the best predictor now trade even more aggressively on their signal (i.e., they make larger bets for a given deviation between the asset price and their forecast of its payoff) because they face less risk (the "aggressiveness effect"). As a result, the asset price is more informative (closer to the asset payoff), which reduces the value of searching for a predictor. When τ^{max} is large enough, the aggressiveness effect dominates and speculators optimally react by adopting a less demanding search policy (i.e., τ^* decreases in equilibrium).

In contrast, a decrease in c unambiguously raises the value of searching for another

Financial Times, August 28, 2017) and supported by recent empirical findings. For instance, Katona et al. (2019) find that combining satellite images of parking lots of U.S. retailers from two distinct data providers improves the accuracy of the forecasts of retailers' quarterly earnings (see also Zhu (2019)). Also, van Binsbergen et al. (2020) find that, with machine learning techniques, one can obtain more precise forecasts of firms' future earnings than analysts' forecasts (they use random forests regressions combining more than 70 accounting variables with analysts' forecasts). Last, Gu et al. (2020) consider 900+ predictors of stock and market returns and find that machine learning techniques (trees and neural networks) considerably increase out-of-sample R^2 of predictive models.

predictor after finding one because it reduces the total expected cost of search without affecting speculators' average trading aggressiveness. Thus, speculators optimally react to a decrease in c by adopting a more demanding search policy (which, in equilibrium, eventually raises their average aggressiveness). The effect of a decrease in α (the "needle in the haystack" effect) is symmetric because it also increases the total expected cost of search without affecting speculators' average trading aggressiveness.

In sum, greater computing power always reduces the difference between the quality of the best and the worst predictor used in equilibrium while data abundance (an increase in τ^{max} or a decrease in α) has the opposite effect (in the case of τ^{max} when it is large enough). These contrasting effects yield several testable implications.

Our first set of predictions is about the informativeness of asset prices for fundamentals. Our model predicts that greater computing power improves price informativeness because it leads speculators to be more demanding for the quality of their predictors. The effect of a push back of the data frontier (due for instance to the availability of a new type of data) on asset price informativeness is more complex. On the one hand, it can lead speculators to be less demanding for the quality, τ^*, of the least satisficing predictor. On the other hand, as it increases the quality of of the best predictor (τ^{max}), it increases the trading aggressiveness of speculators who find the best predictors. The first effect reduces speculators' average trading aggressiveness while the second effect increases it. We find that the second effect always dominates the first in equilibrium. Hence, a push back of the data frontier has a positive effect on price informativeness, even though it can induce some speculators to trade on predictors of lower quality. In contrast, an increase in the severity of the needle in the haystack problem (a decrease in α) always reduces price informativeness.

Our second set of predictions regards the effects of shocks to computing power (e.g., the introduction of cloud computing) or data abundance (e.g., the introduction of new types of data) on the heterogeneity of quantitative asset managers ("quant funds"), measured in various ways. First, we analyze the effects of computing power and data abundance on the cross-sectional distribution of asset managers' trading profits (or equivalently square Sharpe ratios), in particular the mean and the variance of this distribution . Greater computing power raises the average quality of the predictors used in equilibrium and therefore price informativeness. The first effect raises speculators' expected trading profit while the

second reduces it. The former dominates if and only if speculators' cost of exploration, c, is small enough. An increase in the quality of the most informative predictor, τ^{max}, has the same effect for the same reasons. A decrease in α reduces price informativeness and the average quality of predictors used in equilibrium. The second (first) effect dominates when the needle in the haystack problem becomes sufficiently severe (α is low enough). In sum, the model predicts an inverse U-shape relationship between speculators' trading profits (averaged across managers) and (i) data abundance or (ii) computing power. This suggests that progress in information technologies should initially benefit all quant funds until a point where it starts reducing their profits.

In contrast, the model implies that an increase in computing power reduces the variance of asset managers' trading profits while a push back of the data frontier (or an increase in the severity of the needle in the haystack problem) can increase this variance. Indeed, an increase in computing power induces speculators to be more demanding for the quality of their predictors (τ^* increases). Thus, it reduces the dispersion in the quality of predictors, and therefore trading profits, across speculators. On the contrary, data abundance can induce speculators to be less demanding for the quality of their predictors, which increases the cross-sectional dispersion in their performance. The same type of predictions obtain if one considers the heterogeneity of asset managers' investment skills, measured by the predictive power of their holding of the asset for the return of the risky asset (as in Kacperczyk et al. (2014)). This offers another way to test the predictions of the model.

Another way to measure the heterogeneity in asset managers is by the pairwise correlation of their holdings. Intuitively, two managers are less similar when their holdings are less correlated. The model predicts that greater computing power (smaller c) or a push back of the data frontier (greater α) reduce the pairwise correlation in speculators' trades (i.e., increase the heterogeneity of their positions). The reason is that, in equilibrium, speculators optimally trade on the component of their predictors that is orthogonal to the equilibrium price. As c decreases or τ^{max} increases, this component increasingly reflects the noise in speculators' signals because the asset price becomes more informative. As this noise is independent across speculators, speculators' holdings become less correlated when c decreases or τ^{max} increases (an increase in the severity of the needle in the haystack problem has the opposite effect because it reduces price informativeness).

Interestingly, this happens even though speculators may become more similar in terms of the quality of their signal (e.g., in the case of a decrease in c).

2. Contribution to the Literature

Our paper contributes to the literature on informed trading in financial markets when information acquisition is endogenous (see Veldkamp (2011) for a survey). This literature often takes a reduced-form approach to model the cost of acquiring a signal of given precision. For instance, Verrecchia (1982) (and several subsequent papers) assumes that this cost is a convex function of the precision of the signal. The learning technology in our model is different. The relationship between a speculator's total expected cost of obtaining information and the expected precision of her signal is endogenous and micro-founded by an optimal search model. As explained previously, this approach enables us to analyze separately the effects of greater computing power (a decrease in the cost of processing data) and data abundance (an expansion of the search space). To our knowledge, our paper is the first to offer this possibility.

A few other papers have formalized information acquisition as a search problem (Garleanu and Pedersen (2018), Han and Sangiorgi (2018), Banerjee and Breon-Drish (2020)) but they analyze different questions. In Garleanu and Pedersen (2018), investors can invest in passive or active funds and pay a search cost to discover whether an active asset manager is informed or not about a risky asset. In contrast to our model, informed managers have a signal of the same precision obtained as in Grossman and Stiglitz (1980). In Han and Sangiorgi (2018), an agent can draw, with replacement, normally distributed signals from an "urn." Each draw is costly, similar to the cost of exploration in our model. Interestingly, the relationship between the precision of the average signal obtained by the agent (a sufficient statistics for all his signals) and her total investment in drawing signals is convex, which provides a microfoundation for the assumption that information acquisition costs are convex in precision. Our approach differs in many respects. In particular, we jointly solve for the equilibrium of the market for a risky asset and speculators' optimal search for predictors (Han and Sangiorgi (2018) do not apply their model to trading in financial markets). In Banerjee and Breon-Drish (2020), one investor dynamically controls her timing for information acquisition about the payoff of a risky asset. She optimally

alternates between periods in which she searches for information (when the volume of noise trading is high enough) and periods in which she does not. When she searches for information, the investor finds a signal of a given precision according to a Poisson process and starts trading on this signal as soon as she finds it. Banerjee and Breon-Drish (2020) shows that this dynamic model generates predictions different from the standard static model in which the informed investor must decide to acquire a signal before trading.

More broadly, our paper is related to the growing literature on the theoretical effects of new information technologies for the production of financial information (e.g., Abis (2020), Dugast and Foucault (2018), Farboodi and Veldkamp (2019), Milhet (2020) or Huang et al. (2020)). This literature assumes that progress in information technologies reduces the cost of processing information or relaxes investors' attention constraints and explores ramifications of this assumption. Our model accounts for another dimension of this progress, namely data abundance (the expansion of speculators' search space for predictors). We show that the effects of data abundance and the cost of processing data (c in our model) are different and derive several implications that should allow empiricists to test whether these differences matter empirically.

Last, there is a growing literature in financial economics on the risk of false discovery (p-hacking) due to extensive data mining (e.g., Harvey (2017)). Our model does not speak to this issue. We interpret the search for predictors as data mining. However, when a speculator discovers a predictor, there is no uncertainty about the quality of the predictor (see Footnote 9 for more discussion).

3. Model

3.1 Searching predictors

We consider a financial market with a continuum of risk averse (CARA with risk aversion ρ) speculators of unit mass, risk neutral and competitive market makers, and noise traders. Speculators can invest in a risky asset and a risk free asset whose rate of return is normalized to zero. Speculators have no initial endowments in these assets. We interpret speculators in our model as managers of funds who have the infrastructure to process vast amount data and trade on signals extracted from these data ("quantitative funds"). As in Garleanu and Pedersen (2018)'s model of active asset management, we consider the

case in which speculators invest in only one risky asset for simplicity. Our focus is on how speculators discover their trading signals. In Section 6, we endogenize the decision to become speculator (i.e., the fraction of quantitative asset managers).

Figure 1 describes the timing of the model. The payoff of the risky asset, ω, is realized in period 2 and is normally distributed with mean zero and variance σ^2. Speculators search a predictor of the asset payoff in period 0 (the "exploration stage"). Then, in period 1 (the "trading stage"), they observe the realization of their predictor and can trade in the market for the risky asset.

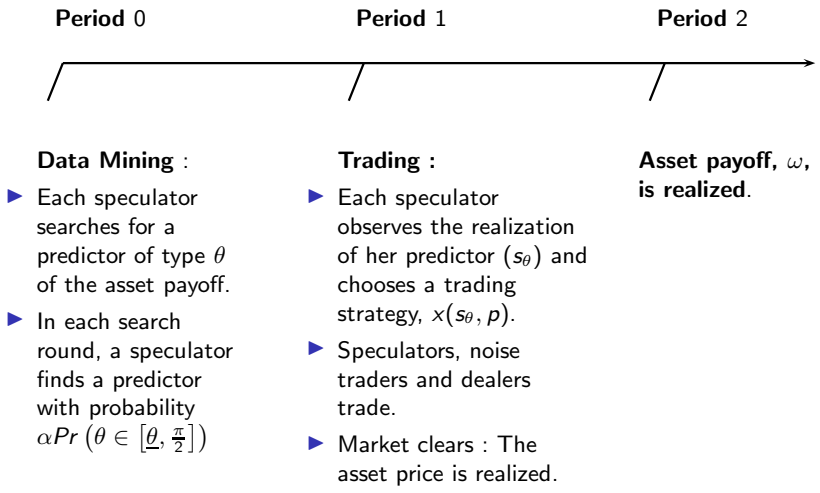


Figure 1: Timing

The exploration stage. In period 0, each speculator i searches for a *predictor* of the asset payoff, ω. There is a continuum of potential predictors. Each predictor, s_θ, is characterized by its type θ and is such that:

$$s_\theta = \cos(\theta)\omega + \sin(\theta)\varepsilon_\theta, \tag{1}$$

where $\theta \in [0, \pi/2]$ and the ε_θs are normally and independently distributed with mean zero and variance σ^2.Moreover, ε_θ is independent from ω. Let $\tau(\theta) \equiv \cos^2(\theta)/\sin^2(\theta) = \cot^2(\theta)$ denote the signal-to-noise ratio for a predictor with type θ. We refer to this ratio as the "quality" of a predictor. The quality of a predictor decreases with its type, θ and varies from zero ($\theta = \frac{\pi}{2}$) to infinity (when θ goes to zero). It is unrelated to the uncertainty about

the asset payoff, σ^2, because $\mathsf{Var}[\varepsilon_\theta] = \mathsf{Var}[\omega] = \sigma^2$.[7] We assume that predictors' types, θs, are distributed according to the cumulative probability distribution $\Phi(.)$ (density $\phi(.)$) on $[0, \pi/2]$.

An alternative, more standard, approach is to assume that speculators can find predictors $\hat{s}_\tau$ specified as $\hat{s}_\tau = \omega + \tau^{-\frac{1}{2}}\varepsilon_\theta$ and use the distribution of τ as a primitive of the model. In either case, τ determines the informativeness of a predictor because $\mathsf{Var}(\omega \mid \hat{s}_\tau) = \mathsf{Var}(\omega \mid s_\theta) = (1+\tau)^{-1}\sigma^2$ (a predictor with a larger τ reduces a speculator's uncertainty about the asset payoff by a larger amount). We show in Section II.B of the online appendix that results with this specification are identical. In particular, if $\tau = \tau(\theta)$, the predictor $\hat{s}_\tau$ is identical to the predictor s_θ in the sense that a speculator behaves exactly in the same way for each predictor. Our approach is just a change in variable, which proves convenient for calculations of various moments required for solving the model. In Section II.B of the online appendix, we also show how one can obtain the probability distribution of $\tau(\theta)$ for a given probability distribution of predictor's type θ.

Speculators discover predictors' types in period 0 via a sequential search process comprising multiple rounds of exploration. Each round costs c and possibly yields a new type of predictor in $[\underline{\theta}, \frac{\pi}{2})$, i.e., speculators cannot find predictors with quality higher than $\tau^{max} \equiv \tau(\underline{\theta})$. More specifically, with probability $\alpha(1-\Phi(\underline{\theta}))$ $(0 < \alpha \leq 1)$, an exploration round is successful and returns a predictor of type θ (picked according to the distribution $\phi(.)$) in $[\underline{\theta}, \frac{\pi}{2})$. Otherwise, it returns a predictor that is just noise.[8] After each exploration round, a speculator can decide (i) to stop searching and to trade in period 1 on the predictor she just found or (ii) to start a new exploration. There is no limit on the number of exploration rounds.

It is worth stressing that speculators observe the realization of their chosen predictor, s_θ, in period 1, *not* in period 0. In period 0, each speculator just chooses the type (quality) of her predictor. A predictor can be viewed as a function (determined, for instance, with linear regressions or machine learning techniques) of variables from different datasets (e.g., accounting data, geolocation data and consumer transactions data) that minimizes the

[7] Without this assumption, the quality of all predictors would, counter-intuitively, increase with uncertainty.

[8] We assume that speculators draw the type of their predictors according to the unconditional distribution of predictors' type ($\phi(.)$) in the interval $[0, \pi/2]$ but that they cannot exploit predictors with a type $\theta < \underline{\theta}$. Alternatively, one can assume that speculators draw the type of their predictors in $[\underline{\theta}, \pi/2]$, conditional on this type being in this interval. We show in the online appendix (Section II.D) that this approach yields the same results.

predictor's average forecasting error in-sample. The speculator then uses the realization of these variables at date 1 to compute the predictor, s_θ, at this date (out-of-sample).[9]

As more datasets become available ("data abundance"), investors can try more diverse variables to predict asset payoffs (even holding the number of variables used to build predictors constant). This evolution has two consequences controlled by parameters $\underline{\theta}$ and α in the model. First, it pushes back the "data frontier", i.e., it improves(at least weakly) the quality of the most informative predictor (the "hidden gold nugget effect.") This dimension of data abundance is controlled by $\underline{\theta}$ in our model: When $\underline{\theta}$ decreases, the quality of the best predictor (the "hidden gold nugget"), τ^{max} improves.

Second, while the number of combinations of variables that one can consider to build predictors becomes very large, the number of combinations that actually yield informative predictors might fall. For instance, there are myriads of ways in which one could combine traffic data in large cities with other data to predict economic growth. However, only a few are likely to be informative and discovering these combinations take time. We refer to this dimension of data abundance as the "needle in the haystack problem."[10] It is controlled by α in our model: As α decreases, a round of exploration is less likely to be successfull.[11]

Finally, parameter c represents the cost of analyzing the predictive power of a specific set of variables (possibily drom different datasets; see the online appendix II.C) to find a predictor. It includes the cost of cleaning and preparing the data for analysis, running sta-

[9] For instance, the predictor could be obtained by running a regression of ω on some variables (see Section II.C in the online appendix for a formalization). In this approach, the R^2 of the regression is a measure of the quality of the predictor. Indeed, the theoretical R^2 of a regression of ω on s_θ (i.e., $1-\mathsf{Var}[\omega \mid s_\theta]/\mathsf{Var}[\omega]$) is equal to $\cos^2(\theta)$. Thus, the higher the quality of a predictor, the higher the R^2 of a regression of the asset payoff on the predictor. In other words, searching for predictors of high quality is the same thing as searching for predictors with high R^2s. Note that, as usual in rational expectations model, we assume that there is no uncertainty on θ, i.e., on the true predictive model relating the payoff of the asset to the predictor. In reality, investors might be uncertain about the true R^2 of a predictive model (e.g., because of too few past observations for past cash-flows relative to the number of variables used to forecast these cash-flows) and learn it over time (see Martin and Nagel (2020)). In our model, this means that speculators would learn about the true θ of a predictor (e.g., after observing an estimate of θ). We leave the analysis of this problem for future work.

[10] Agrawal et al. (2019) discusses a related problem for the generation of new scientific ideas. Specifically, as the space of possible combinations of existing ideas to create new ones enlarges, it becomes more difficult to identify new useful combinations. One can think of the search for predictors at date 0 as a search for new "ideas" to forecast asset payoff. Each new idea is characterized by its forecasting power.

[11] See for instance "*The quant fund investing in humans not algorithms*" (AlphaVille, Financial Times, December 6, 2017), reporting discussions with a manager from TwoSigma noting that: *"Data are noise. Drawing a tradable signal from that noise, meanwhile, takes work, since the signal is continuously evolving [...] Crucially, Duncombe added, there's qualitative data decay going on too. Back in the day, star managers may have had access to far smaller data sets, but the data in hand was of much higher quality."*

tistical softwares to find optimal predictor with the predictive variables, and batckesting trading strategies to assess the economic value of a predictor. These tasks require human capital (e.g., data scientists time) and involve opportunity costs (processing capacity dedicated to these tasks cannot be used for another task). Increase in computing power reduces this cost as it allows to complete an exploration round faster.[12] Importantly, c should not be interpreted as the investments required in infrastructure and datasets to become a speculator (e.g., a quant fund). The cost of these investments is largely fixed and we assume that it has been sunk by speculators before period 0. We consider the decision to pay this cost to be a speculator in Section .

We focus on equilibria in which each speculator follows an optimal stopping rule θ_i^*. That is, speculator i stops searching for new predictors once she finds a predictor with type $\theta \in [\underline{\theta}, \theta^*]$ (a predictor of sufficiently high quality in the feasible range). We denote by $\Lambda(\theta_i^*; \underline{\theta}, \alpha)$ the likelihood of this event for speculator i in a given search round:

$$\Lambda(\theta_i^*; \underline{\theta}, \alpha) \equiv \alpha \Pr(\theta \in [\underline{\theta}, \theta_i^*]) = \alpha \times (\Phi(\theta_i^*) - \Phi(\underline{\theta})) \tag{2}$$

Thus, a decrease in $\underline{\theta}$ raises the likelihood of finding a predictor in a given exploration, holding α constant. This effect captures the idea that while data abundance might reduce the fraction of informative datasets, it increases the chance of finding a good predictor once one has identified an informative dataset.

As the outcome of each exploration is random, the realized number of explorations varies across speculators (even if they use the same stopping rule). Let n_i be the realized number of search rounds for speculator i. This number follows a geometric distribution with parameter $\Lambda(\theta_i^*; \underline{\theta}, \alpha)$. Thus, the expected number of explorations for a given speculator (a measure of her search intensity) is:

$$\mathsf{E}[n_i] = \Lambda(\theta_i^*; \underline{\theta}, \alpha)^{-1}. \tag{3}$$

The trading stage. Trading begins after *all* speculators find a predictor with satisficing quality. At the beginning of period 1, each speculator observes the realization of her

[12] For instance, Anthony Ledford, the chief scientist of MAN AHL (a quantitative fund), writes that *"Strategies based on NLP [...] are also live in client trading. Researching such strategies requires [...] a processor called graphical processing unit (GPU) that can complete the calculations [...] in* 1/30*th of the time taken by [...] a standard computer."* See AI Pioneers in Investment Management, CFA Institute, 2019.

predictor, s_θ and chooses a trading strategy, i.e., a demand schedule, $x_i(s_\theta, p)$, where, p, is the asset price in period 1.

As in Vives (1995), speculators trade with noise traders and risk-neutral market makers. The noise traders' aggregate demand is price-inelastic and denoted by η, where $\eta \sim \mathcal{N}(0, \nu^2)$ (η is independent of ω and errors' in speculators' signals). Market-makers observe investors' aggregate demand, $D(p) = \int x_i(s_\theta, p)di + \eta$ and behave competitively. The equilibrium price, p^* is equal to their expectation of the asset payoff conditional on aggregate demand from noise traders and speculators:

$$p^* = \mathsf{E}\left[\omega \,|D(p^*)\right]. \tag{4}$$

Speculators' objective function. At $t = 2$, the asset pays off and speculator i's final wealth is

$$W_i = x_i(s_\theta, p)(\omega - p) - n_i c. \tag{5}$$

The number of exploration rounds for speculator i, n_i, is independent from the asset payoff, its price, and the realization of the speculator's predictor, s_θ, because n_i is determined in period 0, before the realizations of these variables. Thus, the ex-ante expected utility of a speculator is:

$$\mathsf{E}\left[-\exp(-\rho W_i)\right] = \underbrace{\mathsf{E}\left[-\exp(-\rho(x_i(s_\theta, p)(\omega - p))\right]}_{\text{Expected Utility from Trading}} \times \underbrace{\mathsf{E}\left[\exp(\rho(n_i c))\right]}_{\text{Expected Utility Cost of Exploration}} \tag{6}$$

The first term in this expression represents the ex-ante expected utility that a speculator derives from trading gross of her total exploration cost while the second term represents the expected utility of the total cost paid to find a predictor (we call it the expected utility cost of exploration). The expected utility from trading depends both on the investor's optimal trading strategy ($x_i(s_{\theta,i}, p)$) and her optimal stopping rule (θ_i^*) because this rule determines the distribution of s_θ. The expected utility cost of exploration depends on the speculator's stopping rule, θ_i^*, because it determines the distribution of n_i. In the literature (e.g., Grossman and Stiglitz (1980)), $n_i = 1$ (investors pays a cost and gets one signal of known quality). In our model, n_i is random and its distribution is controlled by the speculator via her stopping rule. Each speculator chooses her stopping rule, θ_i^*, and her trading strategy, $x_i(s_{\theta,i}, p)$, to maximize her ex-ante expected utility.

3.2 Discussion of the modeling choices

In our framework, a new round of exploration does not necessarily yield a predictor of better quality than in a previous round. At the first glance, this assumption may look unrealistic: As speculators accumulate use an increasing number of variables to form their predictors over exploration round, the quality of their predictors in a given round should necessarily be larger than in previous rounds. However, this reasoning assumes that speculators use an increasing number of variables in their predictive model. In reality, asset management firms are likely to constrain their researchers to use a limited number of variables, both to avoid the risk of overfitting and to limit data costs. In Section II.C of the online appendix, we consider such a scenario. We explicitly formalize how speculators build a predictor in each round by using N variables (N can be large; what matters is that it is fixed over all exploration rounds) and show that in this case, the quality of predictors does not necessarily increase from one exploration round to the next (the randomness in the quality of predictors stems from randomness in the predictive power of new variables considered in a given round).

We also assume that if a speculator turns down a predictor, she "forgets" it. We make this assumption because it simplifies the exposition. However, in Section II.A of the online appendix, we relax it. That is, when a speculator stops searching for a predictor, she can trade on the best predictor she found until this moment. Thus, the state variable for a speculator problem is the best predictor she found so far and by definition, the quality of this predictor cannot decline. We show in Section II.A of the online apendix that the results in this case are identical to those obtained in our simpler framework. Indeed, due to the stationarity of speculators' search problem, the optimal stopping rule is identical in both problems.

Last, we assume that all speculators are ex-ante identical. In particular, they have the same exploration costs (c), search set ($\underline{\theta}$), and risk bearing capacity (ρ). In this way, we better highlight how the search for predictors can in itself be an endogenous source of heterogeneity in asset managers' performance. Of course, in reality, there are other sources of heterogeneity between asset managers (e.g., fund size) and these should be controlled for in testing the implications of the model regarding the distribution of asset managers' performance (see Section 6).

4. Equilibrium Data Mining

We focus on symmetric equilibria in which all speculators choose the same stopping rule, θ^*. We proceed as follows. First, we solve for the equilibrium of the trading stage in period 1 taking θ^* as given and we deduce the ex-ante expected utility achieved by speculator i when she chooses a predictor of type θ in period 0. We then observe that a speculator should stop searching when she finds a predictor such that the expected utility of trading on this predictor is larger than or equal to the expected utility she can obtain by launching a new exploration. The optimal stopping rule of each investor, $\theta_i^*(\theta^*)$, is such that this condition holds as an equality (so that the speculator is just indifferent between searching more or stopping). Finally, we pin down θ^* by observing that, in a symmetric equilibrium, each speculator's best response to other speculators' stopping rule, θ^*, must be identical, i.e., $\theta_i^*(\theta^*) = \theta^*$.

Equilibrium of the asset market in period 1. The outcome of the exploration phase is characterized by the distribution of the predictors' types chosen by speculators. Let $\phi^*(\theta; \theta^*; \underline{\theta}, \alpha)$ be this distribution given that speculators follow the stopping rule θ^*:

$$\phi^*(\theta; \theta^*; \underline{\theta}, \alpha) = \frac{\alpha\phi(\theta)}{\Lambda(\theta^*; \underline{\theta}, \alpha)}. \tag{7}$$

We denote the *average* quality of predictors across all speculators in period 1 by $\bar{\tau}(\theta^*, \underline{\theta}, \alpha) \equiv \mathsf{E}\left[\tau(\theta) | \, \underline{\theta} \leq \theta \leq \theta^*\right]$ and we assume that $\phi(.)$ is such that $\bar{\tau}(\theta^*, \underline{\theta}, \alpha)$ is well defined even when $\underline{\theta} = 0$, that is,

A.1: The distribution of predictors' type, $\phi(.)$, is such that for all $\theta^* > 0$, $\bar{\tau}(\theta^*; 0, \alpha)$ exists.[13]

Proposition 1 provides the equilibrium of the asset market in period 1.

Proposition 1. *In period 1, the equilibrium trading strategy of a speculator with type θ is*

$$x^*(s_\theta, p) = \frac{\mathsf{E}[\omega | s_\theta, p] - p}{\rho \, \mathsf{Var}[\omega | s_\theta, p]} = \frac{\tau(\theta)}{\rho\sigma^2}\left(\hat{s}_{\tau(\theta)} - p\right), \tag{8}$$

[13]For some distributions of predictors' type, $\phi(.)$, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$ can diverge because $\tau(\theta)$ goes to infinity when θ goes to zero. Assumption A.1 rules out these distributions.

where $\hat{s}_{\tau(\theta)} = \omega + \tau(\theta)^{-1/2}\varepsilon_\theta$ *and the equilibrium price of the asset is*

$$p^* = \mathsf{E}[\omega | D(p)] = \lambda(\theta^*)\xi. \tag{9}$$

where

$$\xi \equiv \omega + \rho\sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^{-1}\eta, \quad \textit{and} \quad \lambda(\theta^*) \equiv \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2}{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2 + \rho^2\sigma^2\nu^2}, \tag{10}$$

This result extends Proposition 1.1 in Vives (1995) to the case in which speculators have signals of heterogenous precisions (determined by their θ in our model). The predictors s_θ and $\hat{s}_{\tau(\theta)}$ are equivalent from the viewpoint of a speculator because $\hat{s}_{\tau(\theta)} = (\cos(\theta)^{-1})s_\theta$. A speculator's optimal position in the asset is equal to the difference between her (equivalent) predictor and the price of the asset scaled by $\frac{\tau(\theta)}{\rho\sigma^2}$. We refer to this scaling factor as the speculator's aggressiveness. Speculators with predictors of higher quality (larger $\tau(\theta)$) trade more aggressively (take larger positions) on the difference between their predictor and the price of the asset because, conditional on their information, they face less uncertainty.

As in Grossman and Stiglitz (1980), we measure the informativeness of the asset price by $\mathcal{I}(\theta^*; \underline{\theta}, \alpha) = \mathsf{Var}[\omega \mid p^*]^{-1}$. Using Proposition 1, we obtain that

$$\mathcal{I}(\theta^*; \underline{\theta}, \alpha) = \tau_\omega + \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2\tau_\omega^2}{\rho^2\nu^2}, \tag{11}$$

where $\tau_\omega = 1/\sigma^2$ is the precision of speculators' prior about the asset payoff. Thus, the asset price is more informative when the average quality of speculator's predictors, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$, increases. Intuitively, the reason is that speculators' average aggressiveness is greater when the average quality of their predictors is higher. Thus, the total demand for the asset ($D(p)$) is more informative because it becomes more driven by speculators' orders than by noise traders. As a result, the market maker can form a more precise forecast of the asset payoff and the asset price is therefore more informative about this payoff. One implication is that the informativeness of the asset price depends on speculators' search policy θ^*: It is inversely related to θ^* because $\bar{\tau}(\theta^*; \underline{\theta}, \alpha)$ decreases with θ^*. Thus, other things equal, price informativeness is smaller when speculators chooses a less demanding stopping rule for the quality of the predictors on which they trade.

Equilibrium of the exploration phase. Using the characterization of the equilibrium of the asset market, we compute a speculator's expected utility from trading ex-ante, i.e., before observing the realization of her predictor and the equilibrium price, when her predictor has type θ and other speculators follow the stopping rule θ^*. We denote this ex-ante expected utility by $g(\theta, \theta^*)$ and refer to it as the trading value of a predictor with type θ. Formally:

$$g(\theta, \theta^*) \equiv \mathsf{E}\left[-\exp(-\rho(x^*(s_\theta, p^*)(\omega - p^*)) \mid \theta_i = \theta\right]. \tag{12}$$

Lemma 1. *In equilibrium, the trading value of a predictor with type θ is*

$$g(\theta, \theta^*) = -\left(1 + \frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta, p^*] - p^*]}{\mathsf{Var}[\omega|s_\theta, p^*]}\right)^{-\frac{1}{2}} = -\left(1 + \frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*; \underline{\theta}, \alpha)}\right)^{-\frac{1}{2}}. \tag{13}$$

The trading value of a predictor increases with its quality and decreases with the informativeness of the asset price.[14] Thus, it is inversely related to the average quality of predictors used by speculators. Hence, the value of a given predictor for a speculator depends on the search strategy followed by other speculators: It is smaller if other speculators are more demanding for the quality of their predictors (i.e., when θ^* decreases).

Armed with Lemma 1, we can now derive a speculator's optimal stopping rule given that other speculators follow the stopping rule θ^*. Let $\widehat{\theta}_i$ be an arbitrary stopping rule for speculator i. The speculator's continuation utility (the expected utility of launching a new round of exploration) after turning down a predictor is

$$J(\widehat{\theta}_i, \theta^*) = \exp(\rho c)\left(\Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha)\,\mathsf{E}\left[g(\theta, \theta^*)\,\middle|\,\underline{\theta} \le \theta \le \widehat{\theta}_i\right] + (1 - \Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha))J(\widehat{\theta}_i, \theta^*).\right) \tag{14}$$

The first term ($\exp(\rho c)$) in eq.(14) is the expected utility cost of running an additional exploration round. The second term is the likelihood that the next exploration round is successful times the average trading value of a predictor conditional on the type of this predictor being satisficing (in $[\underline{\theta}, \widehat{\theta}_i]$). Finally, the third term is the likelihood that

[14] Observe that $\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta,p^*]-p^*]}{\mathsf{Var}[\omega|s_\theta,p^*]} = \frac{\mathsf{E}[(\mathsf{E}[\omega|s_\theta,p^*]-p^*)^2]}{\mathsf{Var}[\omega|s_\theta,p]}$ because $\mathsf{E}[\omega|s_\theta, p^*] - p^* = 0$. Thus, eq.(13) implies that, $\frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*;\underline{\theta},\alpha)} = \mathsf{E}\left[(\frac{\mathsf{E}(R_\theta|s_\theta)}{\sigma_{R_\theta|s_\theta}})^2\right]$, where $R_\theta = \omega/p^* - 1$ is the excess return of a speculator with type θ (the riskless rate of return is normalized to zero) and $\sigma_{R_\theta|s_\theta}$ is the standard deviation of this return conditional on the observation of s_θ. In other words, $\frac{\tau(\theta)\tau_\omega}{\mathcal{I}(\theta^*;\underline{\theta},\alpha)}$ is the equilibrium value of the expected square Sharpe ratio of a speculator trading on a predictor with type θ.

the next exploration is unsuccessful times the speculator's continuation utility when she turns down a predictor. Solving eq.(14) for $J(\widehat{\theta}_i, \theta^*)$, we obtain

$$J(\widehat{\theta}_i, \theta^*) = \underbrace{\left[\frac{\exp(\rho c)\Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha)}{1 - \exp(\rho c)(1 - \Lambda(\widehat{\theta}_i; \underline{\theta}, \alpha))}\right]}_{\text{Expected Utility Cost from Exploration}} \times \underbrace{\mathsf{E}\left[g(\theta, \theta^*) | \, \underline{\theta} \leq \theta \leq \widehat{\theta}_i\right]}_{\text{Expected Utility from Trading}} \tag{15}$$

Now suppose that speculator i has obtained a predictor with quality θ. If she stops exploring the data at this stage, her expected utility is $g(\theta, \theta^*)$ (her cost of exploration to obtain this predictor is sunk). If instead the speculator decides to launch a new round of exploration, her expected utility is $J(\widehat{\theta}_i, \theta^*)$. Thus, she optimally stops searching for a predictor if $g(\theta, \theta^*) \geq J(\widehat{\theta}_i, \theta^*)$ and keeps searching otherwise. As $g(\theta, \theta^*)$ decreases with θ, the speculator's optimal stopping rule, $\theta_i^*(\theta^*)$, is the value of θ such that she is just indifferent between these two options:

$$g(\theta_i^*, \theta^*) = J(\theta_i^*, \theta^*). \tag{16}$$

In a symmetric equilibrium, $\theta_i^*(\theta^*) = \theta^*$. We deduce that θ^* solves

$$g(\theta^*, \theta^*) = J(\theta^*, \theta^*). \tag{17}$$

Using the expression for $J(., \theta^*)$ in eq.(14), we can equivalently rewrite this equilibrium condition as

$$F(\theta^*) = \exp(-\rho c), \tag{18}$$

where

$$F(\theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta^*} r(\theta, \theta^*)\phi(\theta)d\theta + (1 - \Lambda(\theta^*; \underline{\theta}, \alpha)), \quad \text{for } \theta^* \in \left[\underline{\theta}, \frac{\pi}{2}\right], \tag{19}$$

with

$$r(\theta, \theta^*) \equiv \frac{g(\theta, \theta^*)}{g(\theta^*, \theta^*)} = \left(\frac{\tau(\theta^*)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\tau(\theta)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}\right)^{\frac{1}{2}}, \tag{20}$$

where the second equality in eq.(20) follows from eq.(13). The next proposition shows that there is a unique interior solution (i.e., $\theta^* \in (\underline{\theta}, \frac{\pi}{2})$) to the equilibrium condition (18) when c is small enough.

Proposition 2. *There is a unique symmetric equilibrium of the exploration phase in which all speculators are active (i.e., a unique stopping rule such that $\underline{\theta} \leq \theta^* < \pi/2$ common to*

all speculators) if and only if $F(\pi/2) < \exp(-\rho c) < 1$.

When $\exp(-\rho c) \leq F(\pi/2)$ (i.e., when c is large), the expected utility cost of exploration is larger than expected utility of trading. Hence, a speculator is better off not acquiring information at all if she expects all other speculators to obtain predictors with types in $[\underline{\theta}, \pi/2]$. Thus, if the condition in Proposition 2 is not satisfied, there is no symmetric interior equilibrium. In this case, there exist asymmetric equilibria in which only a fraction of all speculators are active, i.e., search for a predictor and trade (if c is not too large). In these equilibria, active speculators search for a predictor with a stopping rule equal to $\theta^* = \pi/2$ while others are inactive (do not search and do not trade). Moreover, the fraction of active speculators is such that all speculators are indifferent between being active or not. Henceforth,we focus on the case in which the equilibrium search strategy, θ^*, is strictly less than $\frac{\pi}{2}$ (i.e., $F(\pi/2) < \exp(-\rho c) < 1$) because (i) our focus is on what happens when the cost of exploration becomes small and (ii) this shortens the exposition.

5. Data abundance, computing power and speculators' search strategy.

In this section we study how data abundance (a decrease in $\underline{\theta}$ and/or α) and greater computing power (a decrease in c) affect speculators' search strategy in equilibrium, i.e., θ^*. Indeed, their stopping rule determines the range of the quality of predictors used in equilibrium and variations in this range (due to shocks to c or $\underline{\theta}$) generate a host of testable implications that we derive in the next section. We are mainly interested in effects that arise when the cost of search (c) become very small or data become very abundant ($\underline{\theta}$ close to zero) as these are probably the relevant range of parameters given progress in information technologies.

Proposition 3. *A decrease in the cost of exploration, c, always reduces the stopping rule* θ^* *used by speculators in equilibrium (*$\partial\theta^*/\partial c > 0$*) and* θ^* *goes to* $\underline{\theta}$ *when c goes to zero. Thus, greater computing power raises the quality,* $\tau(\theta^*)$*, of the worst predictor used by speculators in equilibrium.*

Holding θ^* constant, a decrease in the per-exploration cost, c, directly reduces the expected utility cost of launching a new exploration after finding a predictor (the first term

in bracket in eq.(15)). Hence, it raises the value of searching for another predictor after finding one (i.e., $J(\theta^*, \theta^*)$). This direct effect induces speculators to be more demanding for the quality of their predictor and therefore works to decrease θ^*. One indirect consequence is that the quality of the average predictor improves and therefore, on average, speculators trade more aggressively on their signal (the "aggressiveness effect"). As a result, price informativeness increases. This indirect effect reduces the expected utility from trading on a satisficing predictor (the second term in bracket in eq.(15)) and therefore dampens the direct positive effect of a decrease in c on the value of searching for a better predictor. However, in equilibrium, the aggressiveness effect can never fully offset the positive direct effect of a decrease in c on the value of searching for a predictor.[15]

We now consider the effect of data abundance on speculators' optimal stopping rule. Remember that data abundance has two consequences in the model: (i) it pushes back the data frontier by raising the quality of the best predictor and (ii) it increases the risk for speculators of using datasets which, after exploration, proves to be useless (the needle in the haystack problem).

Proposition 4.

1. *A decrease in the fraction of informative datasets, α, increases speculators' stopping rule, θ^*, in equilibrium ($\partial\theta^*/\partial\alpha < 0$) and therefore reduces the quality, $\tau(\theta^*)$, of the worst predictor used by speculators in equilibrium.*

2. *When $\underline{\theta}$ is low enough (less than a threshold, $\underline{\theta}^{tr}(c)$), a decrease in $\underline{\theta}$ increases speculators' stopping rule in equilibrium ($\partial\theta^*/\partial\underline{\theta} < 0$ for $\underline{\theta} < \underline{\theta}^{tr}(c)$) and reduces the quality, $\tau(\theta^*)$, of the worst predictor used by speculators in equilibrium.*

When the needle in the haystack problem strengthens (α decreases), speculators become less demanding for the quality of their predictors (Part 1 of Proposition 4). Intuitively, a drop in α increases the expected utility cost of launching a new exploration after finding a predictor (the first term in bracket in eq.(15)) because it reduces the likelihood, Λ, of finding a predictor in a given exploration. Thus, after turning down a predictor, speculators expect to go through a larger number of exploration rounds before finding a satisficing predictor, which increases their total cost of search. For this reason, a decrease

[15]Suppose instead that it does (to be contradicted) and that, as a result, for some values of c, a decline in c raises θ^*. Then, speculators' average aggressivess and therefore price informativeness would fall when c declines. But then the value of searching for a new predictor would increase. A contradiction.

in α has the same effects as an increase in c. Note however that one cannot capture the needle in the haystack problem simply by varying c. Indeed, doing so would require to argue that data abundance and computing power have opposite effects on the same parameter. This approach would prevent a clean separation of the effects of data abundance on the one hand and the effects of greater information processing power on the other hand.

More surprisingly, pushing back the data frontier induces speculators to become *less* demanding for the quality of their predictors when $\underline{\theta}$ becomes small enough. The reason is as follows. Holding θ^* constant, a direct effect of a marginal decrease in $\underline{\theta}$ is to increase the average quality of speculators' predictors. Thus, speculators' average aggressiveness increases and, as a result, price informativeness increases. This effect reduces speculators' ex-ante expected utility of trading and therefore the value of searching for a predictor. The improvement in the best predictor acts as a countervailing force because it raises the expected utility of trading on the best predictor (the hidden gold nugget effect) This effect raises the ex-ante expected utility from trading, holding θ^* constant. However, the aggressiveness effect always dominates the hidden gold nugget effect for $\underline{\theta}$ low enough.

To show this more formally, we differentiate speculators' ex-ante expected utility from trading with respect to $-\underline{\theta}$ (so that we consider a marginal *decrease* in $\underline{\theta}$), holding θ^* constant:

$$\begin{aligned} & -\frac{\partial \mathsf{E}\left[g(\theta,\theta^*)|\,\underline{\theta}\leq\theta\leq\theta^*\right]}{\partial\underline{\theta}} \\ &= \frac{\alpha\phi(\underline{\theta})}{\Lambda(\theta^*;\underline{\theta},\alpha)}\left[\underbrace{g(\underline{\theta},\theta^*)-\mathsf{E}\left[g(\theta,\theta^*)|\,\underline{\theta}\leq\theta\leq\theta^*\right]}_{\text{Hidden Gold Nugget Effect}}\underbrace{-\int_{\underline{\theta}}^{\theta^*}\frac{\partial g(\theta,\theta^*)}{\partial\underline{\theta}}\phi(\theta)d\theta}_{\text{Aggressiveness Effect}}\right] \end{aligned} \tag{21}$$

The first term in bracket is the difference between the expected utility of trading on the best predictor and the ex-ante expected utility of trading. It measures the increase in all speculators' ex-ante expected utility of trading following a marginal decrease in $\underline{\theta}$ due to the improvement of the expected utility of trading on the best predictor. The second term in bracket is the loss in speculators' ex-ante expected utility of trading due to the increase in speculators' aggressiveness that follows an improvement in the quality of the best predictor. When $\underline{\theta}$ goes to zero, the residual risk faced by speculators who obtain the best predictor vanishes (they face less and less uncertainty about the asset

payoff). As a result, their aggressiveness become very large and the asset price becomes increasingly closer to the asset payoff (more informative). Thus, speculators' expected trading profit vanish. For this reason, when $\underline{\theta}$ becomes small enough, the aggressiveness effect dominates and speculators' ex-ante expected utility from trading decreases when $\underline{\theta}$ declines. Consequently, the value of searching for a predictor falls and therefore speculators become less demanding for their predictor (θ^* increases).[16] For sufficiently large value of $\underline{\theta}$, the relative strengths of these effects are reversed: the hidden gold nugget effect dominates the aggressiveness effect and in this case, a small decrease in $\underline{\theta}$ raises the value of searching for a predictor. Thus, speculators adopt a more stringent stopping rule in equilibrium (θ^* decreases when $\underline{\theta}$ decreases). Ultimately, the stopping rule used by speculators, θ^* is a U-shaped function of the data frontier, $\underline{\theta}$ (see Figure 2).

As explained previously, when c declines, speculators' average aggressiveness also increases, which reduces speculators' ex-ante expected utility from trading. However, this effect is never strong enough to offset the positive effect of a reduction in expected search costs on speculators' expected utility. The asymmetry between the effects of a decrease in $\underline{\theta}$ and c stems from the fact that a decrease in the exploration cost, c, never fully dissipates speculators' rents while a decrease in $\underline{\theta}$ does. Indeed, consider the polar case in which c goes to zero. In this case, speculators' optimal stopping rule θ^* goes to $\underline{\theta}$ because speculators search until they find the best predictor when search becomes costless. However, as $\underline{\theta}$ remains strictly larger than zero, the quality of speculators' signal remains bounded and therefore speculators' average aggressiveness does not become very large. As a result, speculators' rents do not vanish when c goes to zero while they do when $\underline{\theta}$ goes to zero.

Figure 2 illustrate Propositions 3 and 4 when $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$. This distribution for predictors' types belongs to a more general family for which we can compute $F(.)$ in closed-form and therefore solve for the equilibrium of the model numerically (see Section III.B in the online appendix). For this family of distributions, $\tau(\theta)$ has a power distribution and Assumption A.1 is satisfied (see the online appendix). As shown by

[16]Pushing back the data frontier has a third effect: It increases the chance of finding a satisficing predictor holding the search strategy, θ^* constant ($\Lambda(\theta^*;\underline{\theta},\alpha)$ increases when $\underline{\theta}$ goes down). This effect reduce the expected number of rounds required to find a predictor and therefore reduces the expected utility cost of searching for a new predictor after rejecting one. Thus, like the hidden gold nugget effect, it works to increase speculators' continuation utility. However, the combined forces of this effect and the hidden gold nugget effect, are not sufficient to offset the negative impact of the aggressiveness effect on speculators' continuation value for $\underline{\theta}$ small.

Figure 2, speculators' optimal stopping rule, θ^*, declines when the exploration cost (c) decreases. In contrast, as explained previously, there is a U-shape relationship between θ^* and the data frontier, so that when $\underline{\theta}$ becomes small, a push back of the data frontier raises θ^*.

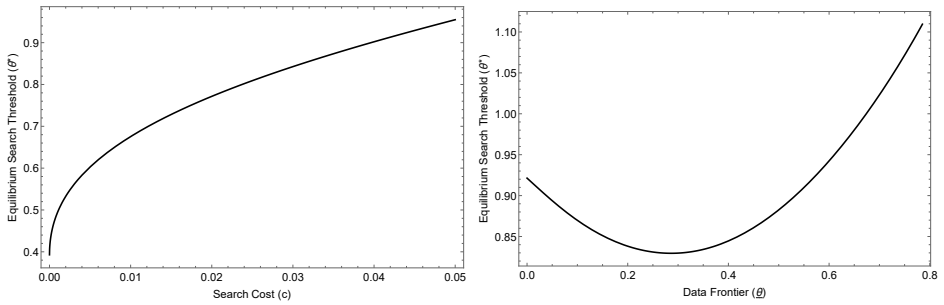


Figure 2: Left-hand-side: Equilibrium search threshold, θ^*, as a function of the search cost, c (other parameter values are $\underline{\theta} = \pi/8, \rho = \sigma^2 = \nu^2 = 1$). Right-hand-side: Equilibrium search threshold, θ^*, as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.03, \rho = \sigma^2 = \nu^2 = 1$).

Proposition 5. *The quality of the worst predictor used in equilibrium, $\tau(\theta^*)$, increases with the volume of noise trading, ν^2, or the volatility of the asset payoff, σ^2.*

Other things equal (in particular θ^*), an increase in the volume of noise trading or the volatility of the asset reduces the informativeness of the equilibrium price. This effect raises the expected value of trading. Thus, the continuation value from searching increases and speculators become therefore more demanding for their predictors (θ^* decreases).

As shocks to computing power or data abundance affect speculators' search strategy (θ^*), they change the average quality of speculators' predictors in equilibrium and therefore the informativeness of the asset price, as stated in the next corollary.

Corollary 1.

1. *In equilibrium, an increase in computing power (a decrease in c) raises the average quality of speculators' predictors and therefore price informativeness.*

2. *In equilibrium, an improvement in the quality of the most informative predictor (a decrease in $\underline{\theta}$) raises the average quality of speculators' predictors and therefore price informativeness.*

3. *In equilibrium, a decrease in the proportion of informative datasets (a decrease in*

α) reduces the average quality of speculators' predictors and therefore price informativeness.

A decrease in computing power induces speculators to be more demanding for the quality of their predictors and thereby raises the average quality of their signals. Thus, price informativeness increases when c declines. A decrease in the fraction of informative datasets (α) has the opposite effect. Holding θ^* constant, a drop in $\underline{\theta}$ increases the average quality of speculators' predictors. However, speculators can react by being less demanding for the quality of their predictors (Proposition 4). This second effect dampens the first but never fully offsets it, so that a push back of the data frontier increase the average quality of speculators' predictors.

Recently, several authors have analyzed the effect of digitization on price efficiency (e.g., Gao and Huang (2019), Zhu (2019), Barbopoulos et al. (2021)), the sensitivity of firms' corporate decisions to prices (e.g., Goldstein et al. (2020))) or the quality of analysts' forecasts at various horizons (Dessaint et al. (2021)). To do so, empiricists use shocks to either the cost of accessing and processing the data (e.g., Gao and Huang (2019), Barbopoulos et al. (2021), or Goldstein et al. (2020)) or the availability of new types of data (e.g., Zhu (2019) or Dessaint et al. (2021)). Corollary 1 (and other implications of the model derived in the next section) suggests that it is important to carefully distinguish between these two types of shocks because they do not necessarily have the same effects. For instance, Corollary 1 predicts that a reduction in the cost of processing information (c) should improve price informativeness. In contrast, an increase in the volume of available data can increase or decrease price informativeness depending on whether or not the negative effect of the needle in the haystack effect on price informativeness dominates the positive effect of pushing back the data frontier.

6. Implications for asset managers' heterogeneity

The contrasting effects of shocks to computing power and data abundance on speculators' search policy imply that shocks to computing power or data abundance should affect the (cross-sectional) dispersion of (a) asset managers' performance and (b) investment skills in opposite directions (see Corollaries 3 and 4 below), even after controlling for other sources of heterogeneity (e.g., differences in size). In contrast, shocks to computing power and the

data frontier should affect their average performance and the correlation of their holdings of the risky asset in the same direction (Corollaries 2 and 5).

6.1 The distribution of asset managers' performance

One way to measure an asset manager's performance is to measure her total dollar return on investment (adjusted for risk). In our model, this corresponds to the trading profit of a speculator equilibrium. The trading profit of a speculator with type θ on her position in the risky asset, denoted $\Pi(s_\theta)$, is

$$\Pi(s_\theta) = x^*(s_\theta, p^*) \times (\omega - p^*), \tag{22}$$

where $x^*(s_\theta, p^*)$ and p^* are given by eq.(8) and eq.(9), respectively. Using eq.(8), we deduce that:

$$x^*(s_\theta, p^*) = \frac{1}{\rho\sigma^2}\left(\tau(\theta)(\omega - p^*) + \tau(\theta)^{1/2}\varepsilon_\theta\right). \tag{23}$$

Thus, using eq.(22), the *expected* trading profit of a speculator with type θ is

$$\bar{\Pi}(\theta) \equiv \mathsf{E}[\Pi(s_\theta)|\theta] = \frac{\tau(\theta)}{\rho\sigma^2}\,\mathsf{Var}[\omega - p^*] = \frac{\tau(\theta)\tau_\omega}{\rho\mathcal{I}(\theta^*, \underline{\theta})}, \tag{24}$$

where the last equality follows from the fact that $p^* = \mathsf{E}(\omega \mid p^*)$ so that $\mathsf{Var}[\omega - p^*] = \mathsf{Var}[\omega \mid p^*] = (\mathcal{I}(\theta^*, \underline{\theta}))^{-1}$ (by definition of $\mathcal{I}(\theta^*, \underline{\theta})$).

Thus, the unconditional expected trading profit of all speculators (the average trading profit across all speculators) is:

$$\mathsf{E}[\bar{\Pi}(\theta)] = \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\rho\sigma^2\mathcal{I}(\theta^*, \underline{\theta})} = \frac{1}{\rho\sigma^2}\left(\frac{\tau_\omega}{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)} + \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\rho^2\nu^2}\right)^{-1}, \tag{25}$$

and the variance of trading profits for speculators (the dispersion of average trading profits across all speculators) is:

$$\mathsf{Var}[\bar{\Pi}(\theta)] = \frac{\mathsf{Var}[\tau(\theta) \mid \underline{\theta} < \theta < \theta^*]}{\sigma^4\rho^2\mathcal{I}^2(\theta^*, \underline{\theta})}. \tag{26}$$

Empirically, $\mathsf{E}[\bar{\Pi}(\theta)]$ and $\mathsf{Var}[\bar{\Pi}(\theta)]$ could be measured by the cross-sectional mean and variance of total trading profits of active funds (for instance in a given quarter). Another possibility is to consider the distribution (across funds) of the squared Sharpe Ratio of

active funds. Indeed, one can show that $\bar{\Pi}(\theta)$ is the expected squared Sharpe ratio of a speculator with type θ, divided by her risk aversion (see Footnote 14). Thus, $\mathsf{E}[\bar{\Pi}(\theta)]$ and $\mathsf{Var}[\bar{\Pi}(\theta)]$ can also be interpreted as the mean and variance of the distribution of squared Sharpe ratios across funds.

An increase in the average quality of predictors ($\bar{\tau}(\theta^*;\underline{\theta},\alpha)$ has an ambiguous effect on speculators' expected profit. On the one hand, as speculators have better predictors on average, they make better investment choices (they are more likely to buy the asset when its return is positive and sell the asset otherwise). On the other, price informativeness increases because speculators trade more aggressively on their signals on average. As shown by eq.(25), the first effect raises speculators' expected profit while the second reduces it. Using eq.(25), we find that the first effect dominates if and only if $\bar{\tau}(\theta^*;\underline{\theta},\alpha) \leq (\tau_\omega\rho^2\nu^2)^{1/2}$. Thus, speculators' average expected profit reaches its maximum for $\bar{\tau}(\theta^*(\underline{\theta},c,\alpha),\underline{\theta},\alpha) = (\tau_\omega\rho^2\nu^2)^{1/2}$ if there are values of $(\underline{\theta},c,\alpha)$ for which this equality holds (we write θ^* as a function of $(\underline{\theta},c,\alpha)$ to emphasize that it depends on the value of these parameters). We deduce the following result.

Corollary 2. Computing power, data abundance and speculators' average performance ($\mathsf{E}[\bar{\Pi}(\theta)]$))

1. *If $\bar{\tau}(\theta^*(\underline{\theta},0,\alpha),\underline{\theta},\alpha) > (\tau_\omega\rho^2\nu^2)^{1/2}$ then speculators' expected profit is a hump shaped function of c, which reaches its maximum for $c=\hat{c}$ (characterized in the proof of the proposition). Otherwise, speculators' expected profit decreases with c and reaches its maximum for $c=0$*

2. *If $\bar{\tau}(\theta^*(0,c,\alpha),0,\alpha) > (\tau_\omega\rho^2\nu^2)^{1/2}$ then speculators' expected profit is a hump shaped function of $\underline{\theta}$, which reaches its maximum for $\underline{\theta}=\hat{\underline{\theta}}$ (characterized in the proof of the proposition). Otherwise, speculators' expected profit decreases with $\underline{\theta}$ and reaches its maximum for $\underline{\theta}=0$.*

3. *If $\bar{\tau}(\theta^*(\underline{\theta},c,1),\underline{\theta},1) > (\tau_\omega\rho^2\nu^2)^{1/2}$ then speculators' expected profit is a hump shaped function of α, which reaches its maximum for $\alpha=\hat{\alpha}$ (characterized in the proof of the proposition). Otherwise, speculators' expected profit increases with α and reaches its maximum for $\alpha=1$*

Corollary 2 generate two important predictions. First, positive shocks to computing power and data abundance should have the same qualitative effects on asset managers'

average performance. Second, these effects become negative when c, $\underline{\theta}$ or α become low, other things equal. This implies that even though data abundance and greater computing power initially increase asset managers' average performance, this evolution should eventually make them worse off (see Figure 3 for a numerical illustration).

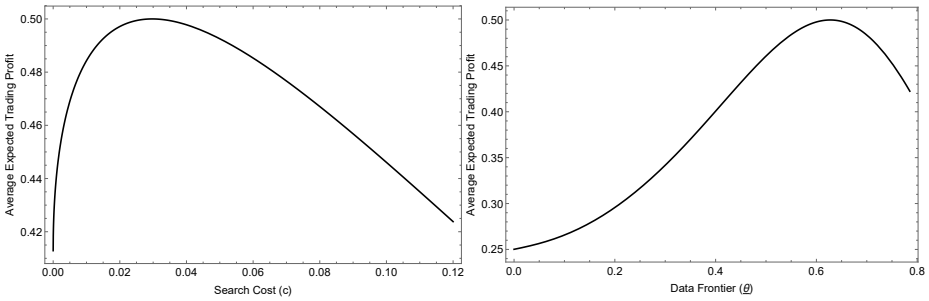


Figure 3: Left: Speculators' expected profits, $\mathsf{E}(\bar{\Pi})$, as a function of the search cost, c (other parameter values are $\underline{\theta} = \pi/5, \rho = \sigma^2 = \nu^2 = 1$). Right: Speculators' expected profits, $\mathsf{E}(\bar{\Pi})$, as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.05, \rho = \sigma^2 = \nu^2 = 1$). In each case $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$.

Corollary 3. Computing power, data abundance and the dispersion of speculators' performance ($\mathsf{Var}[\bar{\Pi}(\theta)]$

1. *Other things equal, for c small enough, the dispersion of speculators' expected trading profit decreases when the cost of processing data goes down (*$d\,\mathsf{Var}[\pi(\theta)]/dc > 0$ *for c sufficiently close to zero).*

2. *Other things equal, for* $\underline{\theta}$ *small enough, the dispersion of speculators' expected profit increases when the data frontier is pushed back (*$d\,\mathsf{Var}[\pi(\theta)]/d\underline{\theta} < 0$ *for* $\underline{\theta}$ *sufficiently close to zero).*

Thus, when c and $\underline{\theta}$ are low enough, a push back the data frontier increases the dispersion of asset managers' performance while reducing the cost of exploration has the opposite effect. The reason is that these parameters have opposite effects on asset managers' managers optimal search strategy. Pushing back the data frontier induces speculators to accept predictors of lower quality (Proposition 4) so that the range of predictors' quality used by speculators ($\tau(\underline{\theta}) - \tau(\theta^*)$) widens. In contrast a decrease in c leads speculators to become more demanding for the quality of their predictors (Proposition 3) so that the range of predictors' quality shrinks. Figure 4 numerically

shows that these results hold for a large range of values for c or $\underline{\theta}$) (the conditions that c and $\underline{\theta}$ are small in Corollary 3 are sufficient but not necessary).

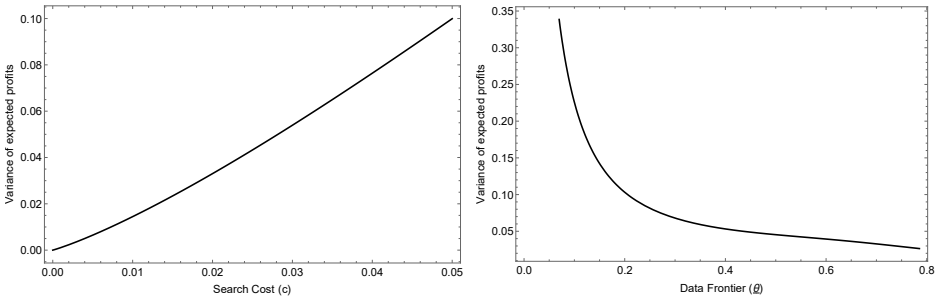


Figure 4: Left: Variance of speculators' expected profits, $\mathsf{Var}[\Pi(\theta)]$, as a function of the search cost, c (other parameter values are $\underline{\theta} = \pi/5, \rho = 1, \sigma^2 = 1, \nu^2 = 1$). Right: Variance of speculators' expected profits as a function of the data frontier, $\underline{\theta}$ (other parameter values are $c = 0.05, \rho = 1, \sigma^2 = 1, \nu^2 = 1$). In each case $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$.

6.2 The distribution of asset managers' investment skills

Kacperczyk and Seru (2007) find empirically that there is heterogeneity in asset managers' investment skills (see their Table I), defined the ability to predict future returns. Our model suggests that this heterogeneity might stem, after controlling for other factors, from serendipity in the search for predictors. If our theory is correct, shocks to computing power or data abundance should affect the dispersion of asset managers' investment skills, as shown below in Corollary 4.

One way to measure an asset manager's investment skills is to study the extent to which he tilts his holdings of a risky asset in the direction of subsequent returns for the asset (see, for instance, Kacperczyk et al. (2014)). One way to do so consists in regressing fund holdings at a given point in time on the subsequent returns on their holdings.[17] If the asset manager has investment skills, the coefficient of this regression should be positive. In our model, the theoretical coefficient, β_θ, of a regression of a speculator's position $(x(s_\theta, p^*))$ on her realized return $(\omega - p^*)$ is

$$\beta_\theta = \frac{\mathsf{Cov}(x(s_\theta, p^*), \omega - p^*)}{\mathsf{Var}[\omega - p^*]} = \frac{\tau(\theta)}{\rho\sigma^2}, \tag{27}$$

[17]Alternatively, one can measure investment skills as in Kacperczyk and Seru (2007). Specifically, Kacperczyk and Seru (2007) measures the precision of asset managers' signals (their "skill") by the sensitivity of their holdings to public information. The higher is this sensitivity, the lower is the precision of a manager's private signals. This would also be the case in a simple extension of our model in which speculators receive a public signal at date 1 in addition to their private signal s_θ.

where the last equality follows from Proposition 1. Holding risk aversion constant, a ranking of speculators based on their investment skill (measured by β_θ) is identical to a ranking based on the (unobservable) quality of their predictors, $\tau(\theta)$.

Let define $\Delta\beta \equiv \frac{\beta(\underline{\theta}))-\beta(\theta^*)}{\beta(\underline{\theta}))} = \frac{\tau(\underline{\theta}))-\tau(\theta^*)}{\tau(\underline{\theta}))}$. Thus, $\Delta\beta$ is the difference between the investment skills of the best and worst speculators (empirically, one could use the difference between the average investment skills of the funds in the top and bottom deciles of the investment skills distribution). This difference is one way to measure the dispersion in asset managers' investment skills. Propositions 3 and 4 yield therefore the following testable implications.

Corollary 4. Computing power, data abundance and the dispersion of speculators' investment skills

1. *Other things equal, the dispersion of speculators' investment skills ($\Delta\beta$) decreases when computing power increases (c decreases).*

2. *Other things equal, for $\underline{\theta}$ low enough (less than $\underline{\theta}^{tr}$), a push back of the data frontier (a decrease in $\underline{\theta}$) increases the dispersion of speculators' investment skills ($\Delta\beta$). The effect of a decrease in α is identical.*

One could also test whether the dispersion of speculators' investment skills ($\Delta\beta$) is reduced in periods of heightened fundamental volatility or noise trading, as implied by Proposition 5. Of course, in testing our predictions, empiricists should control for other factors known to determine asset managers' investment skills (e.g., fund size; see Chen et al. (2004), Ferreira et al. (2012) or Zhu (2018)). Our theory does not say that these factors do not play a role. It just highlights a new source of heterogeneity in investment skills.

6.3 Heterogeneity of asset managers' holdings

Another way to measure the heterogeneity of asset managers is by the correlation of their holdings. The smaller is this correlation, the higher the heterogeneity in asset managers' holdings. Let $\mathsf{Cov}(x(s_{\theta_i}, p^*), x(s_{\theta_j}, p^*))$ be the covariance between the equilibrium holdings of a speculator with type θ_i and a speculator with type θ_j. Using eq.(23) and the fact

that $\mathsf{Var}[\omega - p^*] = (\mathcal{I}(\theta^*, \underline{\theta}))^{-1}$, we obtain:

$$\mathsf{Cov}(x^*(s_{\theta_i}, p^*), x^*(s_{\theta_j}, p^*)) = \frac{\tau(\theta_i)\tau(\theta_j)}{\sigma^4\rho^2}\,\mathsf{Var}[\omega - p^*] = \frac{\tau(\theta_i)\tau(\theta_j)}{\sigma^4\rho^2\mathcal{I}(\theta^*, \underline{\theta})}. \tag{28}$$

Thus, the pairwise correlation between the equilibrium positions of a speculator with type θ_i and a speculator with type θ_j is:

$$\mathsf{Corr}(x^*(s_{\theta_i}, p^*), x^*(s_{\theta_j}, p^*)) = \left(1 + \frac{\mathcal{I}(\theta^*, \underline{\theta})}{\tau(\theta_i)\tau_\omega}\right)^{-\frac{1}{2}} \left(1 + \frac{\mathcal{I}(\theta^*, \underline{\theta})}{\tau(\theta_j)\tau_\omega}\right)^{-\frac{1}{2}} \tag{29}$$

Holding the quality of the predictors used by two speculators constant, their positions become less correlated when price informativeness is higher. The reason is that speculators trade on the component of their forecast of the asset payoff that is orthogonal to the price. This component reflects both the component of the fundamental, ω, that is not reflected into the equilibrium price and the noise in speculators' signal. The higher the first component relative to the second, the higher the pairwise correlation in speculators' positions in the asset. As the price becomes more informative, the first component becomes smaller relative to the noise component and as a result, the pairwise correlation between speculators' positions drops. Using Corollary 1, we deduce the following result.

Corollary 5.

1. *Greater computing power (a decrease in c) or a push back of the data frontier (a reduction in $\underline{\theta}$ reduces the pairwise correlation of speculators' positions.*

2. *A decrease in the fraction of informative datasets, α, increases the pairwise correlation of speculators' positions.*

Testing Corollary 5 requires measuring the pairwise correlation of speculators' positions, holding the quality of their signal constant. One possibility is to estimate the cross-sectional distribution of funds' predictors quality using the method described in Section 6.2 and analyze the effect of shocks to computing power or data abundance on the correlation in the positions of funds in different quantiles of the distribution.

7. Extension: Endogenous entry of speculators

In this section, we relax the assumption of a unit mass of speculators and we endogenize the decision to become a speculator. To this end, we add an entry stage before date $t = 0$. During this entry stage, a unit mass of investors ("asset managers") simultaneously choose one of two options: (i) invest K to become a speculator (a "quantitative fund") or (ii) not invest K and be a discretionary investor. The investment K enables a speculator to access the learning technology described in Section 3.1. In contrast, a discretionary investor has only access to signals of type $\bar{\theta}$ with $\underline{\theta} < \bar{\theta} < \frac{\pi}{2}$. To focus on the choice between being a speculator or a discretionary investor, we assume that discretionary investors obtain their signal at no cost.[18] We interpret K as technological investments in infrastructure and data required to find predictors with quantitative methods. These investments represent a fixed cost that enable asset managers to discover signals using data mining techniques at low cost per exploration round (low c). Discretionary investors do not pay this cost but face a prohibitively high exploration costs, which limits their ability to discover predictors of high quality.[19]

As both speculators and discretionary investors are informed, we refer to them as "informed investors." We denote by μ ($\in [0, 1]$) the fraction of speculators among informed informed investors and by $\theta^*(\mu)$ speculators' optimal stopping time. To simplify the exposition,we focus on the case in which c is small enough so that $\theta^*(1) < \bar{\theta}$ (this condition is always satisfied for c small because $\theta^*(1)$ decreases with c and goes to $\underline{\theta}$ when c goes to zero). This condition guarantees that all speculators trade on signals of better quality than discretionary investors because $\theta^*(\mu)$ increases with μ (Lemma 2). For a fixed μ, the average quality of informed investors' signals is

$$\bar{\tau}(\theta^*, \mu, \underline{\theta}, \bar{\theta}, \alpha) = \mu \, \mathsf{E}\left[\tau(\theta) | \, \underline{\theta} \leq \theta \leq \theta^*\right] + (1 - \mu)\tau(\bar{\theta}). \tag{30}$$

[18] One can extend the analysis to the case in which discretionary investors pay a fixed cost for their signal, as in many other models of information acquisition (e.g., Grossman and Stiglitz (1980). However, in this case, there is a possibility that some investors will decide to be neither speculator, nor discretionary investor because the costs of discretionary investing and K are too high. This possibility increases the number of cases to analyze without adding new economic insights.

[19] Abis (2020) also considers a model with quants and discretionary investors. Our model differs in many ways and does not focus on the same issues. In particular, in contrast to Abis (2020), we do not analyze the type of information (systematic or idiosyncratic collected by each type of investors. Moreover, in this section, we study how data abundance and the cost of computing power affect the proportion of investors becoming speculators(quants) and speculators' search policy. In Abis (2020) the proportion of quants is fixed and data abundance does not play a role.)

For a fixed value of μ, we can proceed as in Section 4 to derive speculators' equilibrium stopping rule with $\bar{\tau}(\theta^*, \mu, \underline{\theta}, \bar{\theta}, \alpha)$ playing the role of $\bar{\tau}(\theta^*, \mu, \underline{\theta}, \bar{\theta}, \alpha)$ in all equations (see Section II.F in the online appendix for details and the proof of the next lemma).

Lemma 2. *Suppose that $\theta^*(1) < \bar{\theta}$. For a fix value of $\mu > 0$, all implications obtained in the case $\mu = 1$ are still qualitatively valid. Moreover, $\theta^*(\mu)$ and price informativeness increase with μ.*

Thus, the implications obtained in Sections 5 and 6 still hold when $0 < \mu < 1$. When μ increases, the average quality of informed investors' signals in equilibrium increases with μ (see eq.(30)), holding θ^* constant. This effect improves price informativeness and therefore reduces speculators' average expected utility from trading. Speculators respond by searching less intensively, which explains why $\theta^*(\mu)$ increases with μ.

Now, we analyze how μ is determined in equilibrium. Let μ^* be the equilibrium value of μ. After paying the cost K, the problem faced by a speculator is identical to that analyzed in Section 4. Thus, gross of the entry cost K, the ex-ante expected utility of a speculator is $J(\theta^*(\mu), \theta^*(\mu)) = g(\theta^*(\mu), \theta^*(\mu))$ (Condition (17)) and, therefore, her expected utility net of the entry cost is $exp(\rho K) g(\theta^*(\mu), \theta^*(\mu))$. If instead an investor becomes a discretionary investor, she obtains an expected utility equal to $g(\bar{\theta}, \theta^*(\mu))$. In an interior equilibrium, i.e., $\mu^* \in (0, 1)$, an investor is just indifferent between being a speculator or a discretionary investor, which requires:

$$exp(\rho K) g(\theta^*(\mu^*), \theta^*(\mu^*) = g(\bar{\theta}, \theta^*(\mu^*)), \tag{31}$$

that is,

$$exp(\rho K)[\frac{g(\theta^*(\mu^*), \theta^*(\mu^*))}{g(\bar{\theta}, \theta^*(\mu^*))}] = 1. \tag{32}$$

The term in bracket on the L.H.S of eq.(32) is the ratio of speculators' expected utility of trading to discretionary investors' expected utility from trading. This ratio is less than 1, i.e., speculators obtain a higher expected utility of trading ($g(\bar{\theta}, \theta^*(\mu^*)) < g(\theta^*(\mu^*), \theta^*(\mu^*)) < 0$) because the quality of speculators's predictors is higher on average. When the mass of speculators, μ increases, the expected utility of all informed investors decreases because price informativeness increases. However, that of speculators decreases faster so that the ratio of speculators' expected utility from trading to discretionary investors' expected utility from trading increases (gets closer to 1), as shown in the proof

of Proposition 6. Thus, if there is an interior equilibrium (a solution to eq.(32)) then it is unique. If instead eq.(32) has no solution, there is no interior equilibrium and, in equilibrium, either no investor becomes a speculator ($\mu^* = 0$) or all investors are speculators ($\mu^* = 1$). The first case arises when K is large enough and the second case when K is low enough.

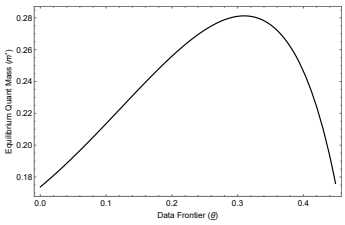
Proposition 6. *The equilibrium mass of speculators, μ^*, is unique. It is equal to 1 when $K \leq K_0$, strictly between 0 and 1 when $K_0 < K < K_1$ and equal to 0 when $K \geq K_1$. When $K_0 < K < K_1$ (i.e., $\mu^* \in (0,1)$), an improvement in computing power (a decrease in c) or a push back of the data frontier (a decrease in $\underline{\theta}$) induces speculators to be more demanding for the quality of their predictors (i.e., $\theta^*(\mu^*)$ decreases when c or $\underline{\theta}$ decrease). Last, a decrease in α induces speculators to be less demanding for the quality of their predictors.*

Thus, when the mass of speculators, μ, adjusts to a change in the data frontier, a push back of the data frontier always leads speculators to be more demanding for the quality of their predictors. In contrast, when μ is fixed, a push back of the data frontier has the opposite effect (see Proposition 4). The reason for this difference is the following. As explained previously, holding μ and θ^* fixed, a decline in $\underline{\theta}$ results in a smaller expected utility from trading for speculators when $\underline{\theta}$ becomes smale enough. When μ is fixed, speculators partially offset this change by searching less intensively for predictors (the intensive margin). When μ is endogenous, there is another margin of adjustment for speculators' expected utility: μ can drop (the extensive margin). Indeed, such a drop reduces the average quality of predictors used by informed traders and price informativeness. As a result, speculators' expected utility from trading increases, which partially offset the negative effect of a drop in $\underline{\theta}$. Thus, when μ can adjust, this is via the extensive margin rather than the intensive margin (the search intensity) that speculators offset the negative effect of a push back of the data frontier on their expected utility from trading.

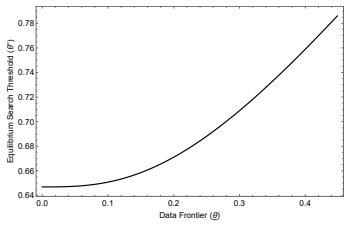
Figure 5 illustrates this point. Panel A shows the evolution of speculators' equilibrium search policy ($\theta^*(\mu^*)$ and the equilibrium mass of speculators as $\underline{\theta}$ declines for various values of $\bar{\theta}$. As it can be seen, the equilibrium mass of speculators is single peaked in $\underline{\theta}$. Thus, there is always a cutoff value for $\underline{\theta}$ after which a decline in $\underline{\theta}$ has a negative effect on μ^*. This effect explains why the effect of $\underline{\theta}$ on $\theta^*(\mu^*)$ is monotonic while it is not when

μ is fixed. Interestingly, as Panel B shows, the effect of a change in c on the equilibrium mass of speculators can be non monotonic as well (this depends on parameter values). Indeed, holding μ constant, speculators search for predictors more intensively when c decreases ($\theta^*(\mu)$ decreases with c; Proposition 3). This response has an ambiguous effect on their relative expected utility from trading. On the one hand, it increases it because it increases their average informational advantage relative to discretionary funds. On the other hand, it increases speculators' aggressiveness and therefore price informativeness, which reduces the gain of informed trading, especially for speculators. This second effect can dominate when c becomes sufficiently small (at this point, a decrease in c triggers a decrease in the equilibrium mass of speculators, μ^*).

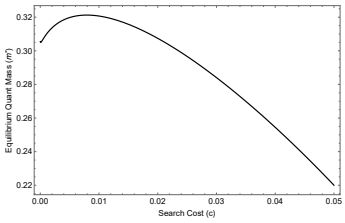
There are two reasons why the case in which μ does not adjust (i.e., the case analyzed in previous sections) is relevant in reality. First, building up the infrastructure and human capital required for quant funds take some time. Thus, the adjustments of μ following shocks to computing power and data abundance cannot be immediate. Thus, our predictions for a fixed μ regarding the effects of $\underline{\theta}$ hold at least in the short term following these shocks (those regarding c and α hold whether μ is fixed or not). Another reason is that investments in infrastructure and technologies (K) are largely sunk costs. Thus, once a speculator has decided to pay these costs, she has no reason to exit (the case in which a decrease in $\underline{\theta}$ triggers a decrease in μ^* means that some speculators exit the market and become discretionary investors). This suggests that once the industry has reached the point at which μ^* reaches its peak, there should be no furher adjustments in μ^* as c or $\underline{\theta}$ keep declining (or very slow adjustments).



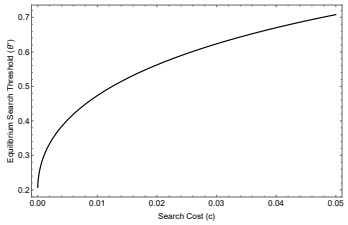
(a) Panel A: Effect of $\underline{\theta}$ on μ^*



(b) Effect of $\underline{\theta}$ on θ^*



(c) Panel B: Effect of c on μ^*



(d) Effect of c on θ^*

Figure 5: Parameter values: $\rho = \sigma^2 = \nu^2 = 1$; $\bar{\theta} = \frac{\pi}{3}$; $K = -log(0.85))$;$c = 0.04$ (upper panel); $\underline{\theta} = \frac{\pi}{8}$ (lower panel). In each case $\phi(\theta) = 3\cos(\theta)\sin^2(\theta)$.

8. Conclusion

Progress in information technologies enable investors to have access to more data (data abundance), both in terms of volume and diversity, and greater computing power, so that they can deploy more powerful techniques to extract information from raw data. In this paper, we propose a new model of information acquisition to analyze separately the effects of these two distinct dimensions of technological progress.

In our model, speculators search (mine data) for predictors via trials and optimally stop searching when they find a predictor with a signal-to-noise ratio larger than an endogenous threshold. As the outcome of speculators' search process is random, speculators discover different predictors. Thus, even though they are homogenous ex-ante, speculators are heterogeneous ex-post in terms of the quality of their predictors, their performance, their holdings etc. In this way, our model generates predictions about the effects of data abundance and computing power on the distribution of asset managers' skills (precisions of their signals), the distribution of their trading profits, or the correlation in their holdings. Moreover, asset price informativeness is determined by speculators' optimal data mining strategy because this strategy determines the average quality of their signals and thereby the informativeness of their aggregate demand.

The main message of our model is that the effects of data abundance and greater computing power are not the same. For instance, greater computing power always induces speculators to be more demanding for the minimal quality of their predictors while this is not necessarily the case for data abundance. As a result, positive shocks to computing power improve and homogenize predictors' quality across speculators and, for this reason, improve price informativeness. In contrast, data abundance can result in a greater dispersion of predictors' quality across speculators and a drop in price informativeness.

References

Abis, Simona, 2020, Man vs machine: Quantitative and discretionary equity management, Working paper, Columbia University.

Agrawal, Ajay, John McHale, and Alexander Oettl, 2019, Finding needles in haystacks: Artifical intelligence and recombinant growth, *in The Economics of Artificial Intelligence, the University of Chicago Press* .

Banerjee, Snehal, and Bradyn Breon-Drish, 2020, Dynamics of research and strategic trading, Working paper, University of California at San Diego.

Barbopoulos, Leonidas, Rui Dai, Talis Putnins, and Anthony Saunders, 2021, Market efficiency in the age of machine learning, Working paper, University of Edinburgh.

Brogaard, Jonathan, and Abalfazl Zareei, 2019, Machine learning and the stock market, Working paper, University of Utah.

Chen, Joseph, Harrison Hong, Ming Huang, and Jeffrey D. Kubik, 2004, Does fund size erode mutual fund performance? the role of liquidity and organization, *American Economic Review* 94.

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard, 2021, Does alternative data affect financial forecasting? the horizon effect, Technical report, Working Paper.

Dugast, Jerome, and Thierry Foucault, 2018, Data abundance and asset price informativeness, *Journal of Financial economics* 130, 367–391.

Farboodi, Maryam, and Laura Veldkamp, 2019, Long run growth of financial technology, *forthcoming American Economic Review* .

Ferreira, Miguel A., Aneel Keswani, António F. Miguel, and Sofia B. Ramos, 2012, The determinants of mutual fund performance: A cross-country study, *Review of Finance* 17.

Gao, Meng, and Jiekun Huang, 2019, Informing the market: The effect of modern information technologies on information production, *The Review of Financial Studies* 1367–1411.

Garleanu, Nicolae, and Lasse Heje Pedersen, 2018, Efficiently inefficient markets for assets and asset management, *Journal of Finance* 78, 1163–1711.

Goldstein, Itay, Shiijie Yang, and Luo Zuo, 2020, The real effects of modern information technologies, *Working paper, NBER* .

Grossman, Sanford, and Joseph Stiglitz, 1980, On the impossibility of informationally efficient markets, *American Economic Review* 70, 393–408.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, 2020, Empirical Asset Pricing via Machine Learning, *The Review of Financial Studies* 33, 2223–2273.

Han, Jungsuk, and Francesco Sangiorgi, 2018, Searching for information, *Journal of Economic Theory* 175, 342–373.

Harvey, Campbell, 2017, The scientific outlook in financial economics, *Journal of Finance* 72, 1399–1440.

Huang, Shyang, Yang Xiong, and Liyan Yang, 2020, Information skills and data sales, *Working paper* .

Kacperczyk, Marcin, and Amit Seru, 2007, Fund managers use of public information: New evidence on managerial skills, *Journal of Finance* 62, 485–528.

Kacperczyk, Marcin, Stijn van Nieuwerburgh, and Laura Veldkamp, 2014, Time-varying fund manager skills, *Journal of Finance* 69, 1455–1483.

Katona, Zsolt, Markus Painter, Panos Patatoukas, and JienYin Zengi, 2019, On the capital market consequences of alternative data: Evidence from outer space, Technical report.

Marenzi, Octavio, 2017, Alternative data: The new frontier in asset management, *Report, Optimas Research* .

Martin, Ian, and Stefan Nagel, 2020, Market efficiency in the age of big data, *Working paper, LSE and University of Chicago* .

Milhet, Roxana, 2020, Financial innovation and the inequality gap, Technical report.

Narang, Rishi, 2013, *Inside the Black Box: A simple guide to quantitative and high-frequency trading* (Wiley, New-York).

van Binsbergen, Jules H., Xiao Han, and Alejandro Lopez-Lira, 2020, Man vs. machine learning: The term structure of earbnings expectations and conditional biases, *Working paper, NBER* .

Veldkamp, Laura, 2011, *Information choice in macroeconomics and finance* (Princeton University Press).

Verrecchia, Robert, 1982, Information acquisition in a noisy rational expectations economy, *Econometrica* 1415–1430.

Vives, Xavier, 1995, Short-term investment and the informational efficiency of the market, *Review of Financial Studies* 8, 125–160.

Zhu, Christina, 2019, Big data as a governance mechanism, *Review of Financial Studies* 32, 2021–2061.

Zhu, Min, 2018, Informative fund size, managerial skill, and investor rationality, *Journal of Financial Economics* 130.

A. Proofs

Proof of Proposition 1. We show that $x^*(s_\theta, p)$ and p^* as given by eq.(8) and eq.(9) form an equilibrium. First, suppose that $x^*(s_\theta, p)$ is given by $x^*(s_\theta, p) = a(\theta)(\hat{s}(\theta) - p)$. In this case, the aggregate demand for the asset is given by:

$$D(p) = \int x^*(s_\theta, p) + \eta = \bar{a}(\omega - p) + \eta, \tag{33}$$

where $\bar{a}$ is the average value of $a(\theta)$ across all speculators ($\bar{a} = E[a(\theta) \mid \theta \in [\underline{\theta}, \theta^*]]$). Hence, observing $D(p)$ (and p) is informationally equivalent to observing $\xi = \omega + \bar{a}^{-1}\eta$. Thus:

$$p^* = \mathsf{E}\left[\omega \mid D(p)\right] = \mathsf{E}[\omega \mid \xi] = \left(\frac{\sigma^2}{\sigma^2 + \bar{a}^{-2}\nu^2}\right)\xi = \left(\frac{\tau_\xi}{\tau_\omega + \tau_\xi}\right)\xi, \tag{34}$$

where $\tau_\xi \equiv \frac{\bar{a}^2}{\nu^2}$ is the precision of ξ as a signal about ω.

Now consider speculators. Using standard calculations in the CARA gaussian framework, we obtain that the optimal demand for the risky asset of a speculator with signal s_θ is:

$$x^*(s_\theta, p) = \frac{\mathsf{E}[\omega | s_\theta, p] - p}{\rho\, \mathsf{Var}[\omega | s_\theta, p]}, \tag{35}$$

As speculators have rational expectations on the price, they anticipate that it is linear in ξ, as in eq.(34). Moreover, let $\hat{s}_\theta \equiv \omega + \tau(\theta)^{-\frac{1}{2}}\epsilon_\theta$, so that $s_\theta = \cos(\theta)\hat{s}_\theta$. Thus,

$$\mathsf{E}[\omega | s_\theta, p] = \mathsf{E}[\omega | \hat{s}_\theta, \xi]. \tag{36}$$

and

$$\mathsf{Var}[\omega | s_\theta, p] = \mathsf{Var}[\omega | \hat{s}_\theta, \xi]. \tag{37}$$

Note that the precision of $\hat{s}_\theta$ is $\tau(\theta)\tau_\omega$. Thus, as all variables are normally distributed and ϵ_θ and η (the noises in $\hat{s}_\theta$ and ξ) are independent, standard calculations yield:

$$\mathsf{E}[\omega | \hat{s}_\theta, \xi] = \frac{\tau(\theta)\tau_\omega \hat{s}_\theta + \tau_\xi \xi}{\tau_\omega + \tau(\theta)\tau_\omega + \tau_\xi}. \tag{38}$$

and

$$\mathsf{Var}[\omega | s_\theta, p] = \frac{1}{\tau_\omega + \tau(\theta)\tau_\omega + \tau_\xi}. \tag{39}$$

Thus, we can rewrite eq.(35) as:

$$x^*(s_\theta, p) = \frac{\tau(\theta)\tau_\omega \hat{s}_\theta + \tau_\xi \xi - (\tau_\omega + \tau(\theta)\tau_\omega + \tau_\xi)p}{\rho}, \tag{40}$$

Using the fact that $p = \frac{\tau_\xi}{\tau_\omega + \tau_\xi}\xi$ we deduce that:

$$x^*(s_\theta, p) = \frac{\tau(\theta)\tau_\omega}{\rho}(\hat{s}_\theta - p) = \frac{\tau(\theta)}{\rho\sigma^2}(\hat{s}_\theta - p). \tag{41}$$

Thus, $x^*(s_\theta, p)$ is as conjectured (and as in eq.(8)) if and only if $a(\theta) = \frac{\tau(\theta)}{\rho\sigma^2}$. If follows that $\bar{a} = \frac{\bar{\tau}(\theta)}{\rho\sigma^2}$. Eq.(9) and eq.(10) in the text immediately follow from substituting this expression for $\bar{a}$ in eq.(34).

In sum we have shown that (i) if dealers expect speculators to follow the trading strategy $x^*(s_\theta, p)$ given by eq.(8) then they set a price given by eq.(9) and (ii) if dealers set a price given by eq.(9) then speculators follow the trading strategy $x^*(s_\theta, p)$ given by eq.(8). Thus, eq.(8) and eq.(9) form an equilibrium. More generally, it is possible to show that this is the unique equilibrium in which speculators' trading strategy is a linear function of their signal and the price.

Proof of Lemma 1. Conditional on the realization of the price at date 1 and her signal, s_θ, the expected utility of trading for an investor given her optimal trading strategy is:

$$\begin{aligned} &\mathsf{E}[-\exp(-\rho x^*(s_\theta, p)(\omega - p)) \mid s_\theta, p] = \\ &-\mathsf{E}\left[\exp\left(-\rho\left(x^*(s_\theta, p)(\mathsf{E}[\omega \mid s_\theta, p] - p) - \frac{\rho(x^*(s_\theta, p))^2}{2}\mathsf{Var}[\omega \mid s_\theta, p]\right)\right)\right]. \end{aligned} \tag{42}$$

Substituting $x^*(s_\theta, p)$ by its expression in eq.(35), we deduce that:

$$\mathsf{E}\left[-\exp(-\rho x^*(s_\theta, p)(\omega - p)) \mid s_\theta, p\right] = -\exp\left(-\frac{(\mathsf{E}[\omega|s_\theta, p] - p)^2}{2\,\mathsf{Var}[\omega|s_\theta, p]}\right) \tag{43}$$

Thus:

$$g(\theta, \theta^*) = -\mathsf{E}\left[\exp\left(-\frac{(\mathsf{E}[\omega|s_\theta, p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|s_\theta, p^*]}\right)\right]. \tag{44}$$

For a normally distributed variable Z with mean 0 and variance σ_Z^2, $\mathsf{E}[\exp(-Z^2)] = (1 + 2\sigma_Z^2)^{-1/2}$. As $\mathsf{E}[\omega|s_\theta, p] - p$, is normally distributed with mean zero, defining $Z =$

$\mathsf{E}[\omega|s_\theta, p] - p$, we deduce that:

$$g(\theta, \theta^*) = -\left(1 + \frac{\mathsf{Var}\left[\mathsf{E}[\omega|s_\theta, p^*] - p\right]}{\mathsf{Var}[\omega|s_\theta, p^*]}\right)^{-1/2} \tag{45}$$

Observe that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta, p^*] - p^*]}{\mathsf{Var}[\omega|s_\theta, p^*]} = \rho^2 \,\mathsf{Var}[\omega|s_\theta, p^*]\,\mathsf{Var}[x^*(s_\theta, p^*)]. \tag{46}$$

Now using the expression for $x^*(s_\theta, p^*)$ in eq.(41), we obtain that:

$$\mathsf{Var}[x^*(s_\theta, p^*)] = \frac{\tau(\theta)^2\tau_\omega^2}{\rho^2}[\mathsf{Var}(\hat{s}_\theta) + \mathsf{Var}(p^*) - 2\,\mathsf{Cov}(\hat{s}_\theta, p^*)]. \tag{47}$$

Using the expression for p^* in eq(34) and the fact that $\hat{s}_\theta = \omega + \tau(\theta)^{-\frac{1}{2}}\epsilon_\theta$, we obtain after some algebra that:

$$\mathsf{Var}[x^*(s_\theta, p^*)] = \frac{\tau(\theta)\tau_\omega(\tau_\omega + \tau_\omega\tau(\theta) + \tau_\xi)}{\rho^2(\tau_\omega + \tau_\xi)}. \tag{48}$$

Thus, using the expression for $\mathsf{Var}[\omega|s_\theta, p^*]$ in eq.(39), we deduce that:

$$\mathsf{Var}[x^*(s_\theta, p^*] = \frac{\tau(\theta)\tau_\omega}{\rho^2(\tau_\omega + \tau_\xi)\,\mathsf{Var}[\omega|s_\theta, p^*]}. \tag{49}$$

Hence, using eq.(46) and the fact that $\tau_\xi = \frac{\bar{\tau}(\theta^*;\underline{\theta},\alpha)^2\tau_\omega^2}{\rho^2\nu^2}$, we deduce that:

$$\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\theta, p] - p]}{\mathsf{Var}[\omega|s_\theta, p]} = \frac{\tau_\omega\tau(\theta)}{\tau_\omega + \frac{(\tau_\omega\bar{\tau}(\theta^*;\underline{\theta},\alpha))^2}{\rho^2\nu^2}}. \tag{50}$$

This yields the expression for $g(\theta, \theta^*)$.

Proof of Proposition 2. The derivative of $F(\theta^*)$ is

$$\frac{\partial F}{\partial \theta^*} = \alpha \int_{\underline{\theta}}^{\theta^*} \frac{\partial r(\theta, \theta^*)}{\partial \theta^*}\phi(\theta)d\theta, \tag{51}$$

where $r(\theta, \theta^*)$ is defined in eq.(20). As θ^* increases, both $\tau(\theta^*)$ and $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ decreases. We deduce that $r(\theta, \theta^*)$ decreases in θ^*. Thus, $\frac{\partial F}{\partial \theta^*} < 0$. Moreover, we have (i) $F(\underline{\theta}) = 1$, (ii) $0 < F(\pi/2) < 1$ and (iii) $\exp(-\rho c) < 1$ (since $c > 0$). Thus, there is a unique solution to the condition $F(\theta^*) = \exp(-\rho c)$ and this solution is in $(\underline{\theta}, \pi/2)$ if and only if

$F(\pi/2) \leq \exp(-\rho c) < 1$.

Proof of Proposition 3. In equilibrium, $F(\theta^*) = \exp(-\rho c)$. The R.H.S of this condition decreases with c and $F(.)$ decreases in θ^* (see the proof of Proposition 2). We deduce that θ^* increase in c. Moreover when c goes to zero, the R.H.S of the equilibrium condition goes to 1. This implies that $F(\theta^*)$ goes to 1 as well, which (by continuity of $F(.)$) is possible only if θ^* goes to $\underline{\theta}$ (as $F(\underline{\theta}) = 1$).

Proof of Proposition 4.

Part 1. It directly follows from eq.(19) that $\frac{\partial F}{\partial \alpha} = -\int_{\underline{\theta}}^{\theta^*}(1 - r(\theta, \theta^*)\phi(\theta)d\theta) < 0$, since $r < 1$. Thus, $F(\theta^*)$ decreases in α. As $F(.)$ also decreases in θ^* and, in equilibrium, $F(\theta^*) = \exp(-\rho c)$, it immediately follows that θ^* increases in α, as claimed in the first part of the proposition.

Part 2. Remember that $\mathcal{I}(\theta^*; \underline{\theta}, \alpha) = \tau_\omega + \frac{\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2 \tau_\omega^2}{\rho^2 \nu^2}$. Thus, we can rewrite $r(\theta, \theta^*)$ given in eq.(20) as:

$$r(\theta, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta^*, \theta^*)} = \left(\frac{\rho^2\sigma^2\nu^2\tau(\theta^*) + \rho^2\sigma^2\nu^2 + \bar{\tau}^2(\theta^*; \underline{\theta}, \alpha)}{\rho^2\sigma^2\nu^2\tau(\theta) + \rho^2\sigma^2\nu^2 + \bar{\tau}^2(\theta^*; \underline{\theta}, \alpha)} \right)^{\frac{1}{2}}. \tag{52}$$

The ratio $(a + x)/(b + x)$ increases with x iff $a < b$. Thus, as $\tau(\theta) > \tau(\theta^*)$, the sign of $\frac{\partial r}{\partial \underline{\theta}}$ is the same as the sign of $\frac{\partial \bar{\tau}}{\partial \underline{\theta}}$ because $\tau(\theta) > \tau(\theta^*)$. We obtain:

$$\frac{\partial \bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} = -\phi^*(\underline{\theta}) \left(\tau(\underline{\theta}) - \bar{\tau}(\theta^*; \underline{\theta}, \alpha)\right) < 0, \tag{53}$$

where the last inequality follows from the fact $\tau(\theta)$ decreases with θ. Thus, $\frac{\partial r}{\partial \underline{\theta}} < 0$.

We deduce from the expression for $\frac{\partial r}{\partial \underline{\theta}}$ that $r(\theta, \theta^*)$ decreases with $\underline{\theta}$ ($\frac{\partial r}{\partial \underline{\theta}} < 0$). Using the expression for $F(.)$ in eq.(19), we deduce that:

$$\frac{\partial F}{\partial \underline{\theta}} = \underbrace{\alpha\phi(\underline{\theta})(1 - r(\underline{\theta}, \theta^*))}_{>0} + \alpha \int_{\underline{\theta}}^{\theta^*} \underbrace{\frac{\partial r}{\partial \underline{\theta}}}_{<0} \phi(\theta)d\theta. \tag{54}$$

Thus, the effect of $\underline{\theta}$ on $F(.)$ and therefore the equilibrium stopping rule θ^* is ambiguous. We now show that this effect becomes negative when $\underline{\theta}$ is close enough to zero. To see this, observe that eq.(54) implies that:

$$\frac{\partial F}{\partial \underline{\theta}} < \alpha\phi(\underline{\theta}) \left(1 + \frac{\int_{\underline{\theta}}^{\theta^*} \frac{\partial r}{\partial \underline{\theta}} \phi(\theta)d\theta}{\phi(\underline{\theta})}\right) \tag{55}$$

We show in Section 4 of the internet appendix that $\frac{\int_{\underline{\theta}}^{\theta^*} \frac{\partial r}{\partial \underline{\theta}} \phi(\theta) d\theta}{\phi(\underline{\theta})}$ goes to $-\infty$ when $\underline{\theta}$ goes to zero. Thus, $\frac{\partial F}{\partial \underline{\theta}} < 0$ for $\underline{\theta}$ small enough. Let $\underline{\theta}^{tr}$ be the smallest value of $\underline{\theta}$ such that $\frac{\partial F}{\partial \underline{\theta}} < 0$. As in equilibrium, $F(\theta^*) = \exp(-\rho c)$ and $F(.)$ decreases in θ^*, it follows that θ^* increases in $\underline{\theta}$ when $\underline{\theta} < \underline{\theta}^{tr}$, as claimed in the second part of the proposition.

Proof of Proposition 5. It follows from direct inspection of the expression for $r(\theta, \theta^*)$ given in eq.(52) that $r(\theta, \theta^*)$ decreases with σ^2, and ν^2 because $\tau(\theta) > \tau(\theta^*)$. Thus, from eq.(19), we deduce that $F(\theta^*)$ decreases with σ^2, and ν^2. It follows from this observation, the fact $F(\theta^*)$ decreases with θ^* and the equilibrium condition $F(\theta^*) = \exp(-\rho c)$ that θ^* decreases with σ^2 and ν^2.

Proof of Corollary 1.

Part 1. Greater computing power induces speculators to be more demanding for the quality of their predictors (to put more effort in the search of good predictors) because it reduces the cost of exploring new data to obtain a predictor (see Proposition 3). Thus, speculators obtain signals of higher quality on average. Hence, on average, they trade more aggressively on their signals, their aggregate demand for an asset becomes more informative and, for this reason, price informativeness increases (see eq.(11)).

Part 2. When a decrease in $\underline{\theta}$ reduces θ^*, it is clear that it raises the average quality of predictors and therefore price informativeness. Now consider the other possible case, i.e., the case in which a decrease in $\underline{\theta}$ increases θ^*. We know that this possibility arises when $\underline{\theta}$ is low enough (see Proposition 4). We prove below, by contradiction, that price informativeness, $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$, is also inversely related to $\underline{\theta}$ in this case.

Suppose (to be contradicted) that there is a value of $\underline{\theta}$ such that when $\frac{\partial \theta^*}{\partial \underline{\theta}} < 0$ then $\frac{\partial \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} > 0$. Let $L(\theta_i^*, \theta^*)$ be:

$$L(\theta_i^*, \theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta_i^*} \frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} \phi(\theta) d\theta + 1 - \alpha \int_{\underline{\theta}}^{\theta_i^*} \phi(\theta) d\theta. \tag{56}$$

Function L is decreasing with θ_i^* because:

$$\frac{\partial L}{\partial \theta_i^*} = \alpha \int_{\underline{\theta}}^{\theta_i^*} \frac{\partial}{\partial \theta_i^*} \left(\frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} \right) \phi(\theta) d\theta < 0. \tag{57}$$

Now, using the expression for $J(.)$ given in eq.(15), we can rewrite the indifference con-

dition (16) as:

$$L(\theta_i^*, \theta^*) = \exp(-\rho c). \tag{58}$$

Moreover: $L(\underline{\theta}, \theta^*) = 1$ and $0 < L(\pi/2, \theta^*) < 1$. Thus, as $L(\theta_i^*, \theta^*)$ decreases in θ_i^* , eq.(56) has a unique solution $\theta_i^*(\theta^*)$ when c is small enough. This solution defines the best response of a speculator when other speculators choose the stopping rule θ^*.

Next, for $\theta_i^* \geq \theta \geq \underline{\theta}$, define

$$l(\theta, \theta_i^*, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta_i^*, \theta^*)} = \left(\frac{\rho^2\sigma^2\nu^2\tau(\theta_i^*) + \rho^2\nu^2 + \sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2}{\rho^2\sigma^2\nu^2\tau(\theta) + \rho^2\nu^2 + \sigma^2\bar{\tau}(\theta^*; \underline{\theta}, \alpha)^2} \right)^{\frac{1}{2}} = \left(\frac{\tau(\theta_i^*)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\tau(\theta)\tau_\omega + \mathcal{I}(\theta^*; \underline{\theta}, \alpha)} \right)^{\frac{1}{2}}. \tag{59}$$

Clearly, $l(\theta, \theta_i^*, \theta^*)$ increases when $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ increases. Thus, if $\frac{\partial \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} < 0$, then $\frac{\partial l(\theta, \theta_i^*, \theta^*)}{\partial \underline{\theta}} > 0$ since $\underline{\theta}$ affects $l(\theta, \theta_i^*, \theta^*)$ only through its effect on price informativeness. This implies that:

$$\frac{\partial l}{\partial \underline{\theta}} + \frac{\partial l}{\partial \theta^*} \frac{\partial \theta^*}{\partial \underline{\theta}} > 0. \tag{60}$$

As:

$$L(\theta_i^*, \theta^*) \equiv \alpha \int_{\underline{\theta}}^{\theta_i^*} l(\theta, \theta_i^*, \theta^*) + 1 - \alpha \int_{\underline{\theta}}^{\theta_i^*} \phi(\theta) d\theta, \tag{61}$$

we deduce that:

$$\frac{dL}{d\underline{\theta}} = \frac{\partial L}{\partial \underline{\theta}} + \frac{\partial L}{\partial \theta^*} \frac{\partial \theta^*}{\partial \underline{\theta}} = \underbrace{\alpha\phi(\underline{\theta})(1 - l(\underline{\theta}, \theta_i^*, \theta^*))}_{>0} + \alpha \int_{\underline{\theta}}^{\theta_i^*} \left(\frac{\partial l}{\partial \underline{\theta}} + \frac{\partial l}{\partial \theta^*} \frac{\partial \theta^*}{\partial \underline{\theta}} \right) \phi(\theta) d\theta. \tag{62}$$

Eq.(60) implies that the second term is also positive. Thus, $\frac{dL}{d\underline{\theta}} > 0$. Thus, a decrease in $\underline{\theta}$ results in a smaller value of L, holding θ_i^* constant. As $\partial L / \partial \theta_i^* < 0$ and $L(\theta_i^*, \theta^*) = \exp(-\rho c)$, it follows that in this case θ_i^* increases with $\underline{\theta}$. As, in equilibrium, $\theta_i^* = \theta^*$, this also implies that $\frac{\partial \theta^*}{\partial \underline{\theta}} > 0$. A contradiction with our starting hypothesis. We deduce that when $\frac{\partial \theta^*}{\partial \underline{\theta}} < 0$ then $\frac{\partial \mathcal{I}(\theta^*; \underline{\theta}, \alpha)}{\partial \underline{\theta}} < 0$. Thus, for all values of $\underline{\theta}$, a decrease in $\underline{\theta}$ improves price informativeness.

Part 3. By definition, $\bar{\tau}(\theta^*; \underline{\theta}, \alpha) = \int_{\underline{\theta}}^{\theta^*} \tau(\theta)\phi^*(\theta)\theta$. Using the definition of $\phi^*(\theta)$, we deduce that $\frac{\partial \bar{\tau}(\theta^*; \underline{\theta}, \alpha)}{\partial \alpha} = \frac{\partial \theta^*}{\partial \alpha}(\phi^*(\theta^*)(\tau(\theta^*) - \bar{\tau}(\theta^*; \underline{\theta}, \alpha)) > 0$, where the last inequality follows from the fact that $\tau(\theta)$ decreases with θ and $\frac{\partial \theta^*}{\partial \alpha} < 0$ (see Proposition 4). Hence, price informativeness increases with α because (i) $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$ increases with $\bar{\tau}(\theta^*)$ and (ii) depends on α only through $\bar{\tau}(\theta^*)$ (see eq.(11). This proves the second part of the proposition.

Proof of Corollary 2. Consider the effect of $\underline{\theta}$ on speculators' expected profits. We know from Corollary 1 that $\bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha)$ decreases with $\underline{\theta}$. Moreover, $\lim_{\underline{\theta} \to \frac{\pi}{2}} \bar{\tau}(\theta^*(\underline{\theta}, c, \alpha), \underline{\theta}, \alpha) = \tau(\frac{\pi}{2}) = 0$. Thus, if $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) > (\tau_\omega \rho^2 \nu^2)^{1/2}$, there is a unique value of θ, denoted $\hat{\theta}$, such that $\bar{\tau}(\theta^*(\hat{\theta}, c, \alpha), \hat{\theta}, \alpha) = (\tau_\omega \rho^2 \nu^2)^{1/2}$. Consequently, when $\underline{\theta}$ varies, holding other parameters constant, speculators' expected profit reaches its maximum for $\bar{\tau}(\theta^*, \hat{\theta}, \alpha) = \tau_\omega \rho^2 \nu^2$. If instead, $\bar{\tau}(\theta^*(0, c, \alpha), 0, \alpha) \leq \tau_\omega \rho^2 \nu^2$, then speculators' expected profit always increases as $\underline{\theta}$ decreases. This proves Part 2 of Corollary 2. The proofs of Parts 1 and 3 are similar and therefore omitted for brevity. In these cases, one obtains that $\hat{c}$ and $\hat{\alpha}$ are the unqiue solutions of, respectively, $\bar{\tau}(\theta^*(\underline{\theta}, \hat{c}, \alpha), \underline{\theta}, \alpha) = (\tau_\omega \rho^2 \nu^2)^{1/2}$ and $\bar{\tau}(\theta^*(\underline{\theta}, c, \hat{\alpha}), \underline{\theta}, \hat{\alpha}) = (\tau_\omega \rho^2 \nu^2)^{1/2}$.

Proof of Corollary 3.

Part 1. For a given $\underline{\theta}$, when $c = 0$ we have $\theta^* = \underline{\theta}$ and therefore $\mathsf{Var}[\Pi(\theta)] = 0$, and when $c > 0$, $\theta^* > \underline{\theta}$ and therefore $\mathsf{Var}[\Pi(\theta)] > 0$. Hence, it must be the case that $\mathsf{Var}[\pi(\theta)]$ is strictly increasing with c, for c close enough to 0.

Part 2. In order to analyze the effect of $\underline{\theta}$, it is useful to rewrite $\mathsf{Var}[\Pi(\theta)]$ as follows (using eq.(26) and the definition of $\mathcal{I}(\theta^*; \underline{\theta}, \alpha)$):

$$\mathsf{Var}[\Pi(\theta)] = \frac{\rho^2 \sigma^4 \nu^4 \left(m_2(\theta^*, \underline{\theta}, \alpha) - \bar{\tau}(\theta^*, \underline{\theta}, \alpha)^2 \right)}{(\bar{\tau}(\theta^*, \underline{\theta}, \alpha)^2 + \rho^2 \sigma^2 \nu^2)^2}. \tag{63}$$

where $m_2(\theta^*, \underline{\theta}, \alpha) \equiv \mathsf{E}\left[\cot^4(\theta) | \, \underline{\theta} \leq \theta \leq \theta^*\right]$ is the second order moment of the variable $\tau(\theta)$ (the distribution of the quality of speculators' predictors). The first moment of this distribution is $\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})$. For a given search cost c, we must distinguish two cases. First, if the second moment of the distribution for the variable $\tau(\theta)$ diverges when $\underline{\theta}$ goes to zero (that is, $\lim_{\underline{\theta} \to 0} m_2(\theta^*, \underline{\theta}, \alpha) = +\infty$,), then we also have $\lim_{\underline{\theta} \to 0} \mathsf{Var}[\Pi(\theta)] = +\infty$. Thus, $\mathsf{Var}[\pi(\theta)]$ is strictly decreasing with $\underline{\theta}$, for $\underline{\theta}$ close enough to 0.

If the second moment of the distribution for the variable $\tau(\theta)$ converges when $\underline{\theta}$ goes to zero, the analysis is more complex.[20] Indeed, as shown below, both the second and the first moments of the distribution for $\tau(\theta)$ decreases with $\underline{\theta}$. If the effect on the second moment dominates then $\mathsf{Var}[\Pi(\theta)]$ decreases with $\underline{\theta}$ while if the effect on the first moment dominates then $\mathsf{Var}[\Pi(\theta)]$ increases with $\underline{\theta}$ (see eq.(63)). We show below that for

[20]Notice first that $m_2(\theta^*, \underline{\theta}, \alpha) < \infty$ (the second moment converges) implies that $\phi(\theta) \cot^4(\theta)$ can be integrated in 0. Locally around $\theta = 0$, since $\cot(\theta) \sim \sin^{-1}(\theta) \sim \theta^{-1}$, we have $\phi(\theta) \cot^4(\theta) \sim \phi(\theta) \cot^2(\theta) \theta^{-2}$ As θ^{-2} cannot be integrated in 0, it must be the case $\lim_{\underline{\theta} \to 0} \phi(\theta) \cot^2(\theta) = 0$. This is a necessary condition so that $\phi(\theta) \cot^4(\theta)$ can be integrated.

$\underline{\theta}$ sufficiently close to zero the first effect dominates.

We have:

$$\frac{d\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{d\underline{\theta}} = \frac{\partial \bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{\partial \underline{\theta}} + \frac{\partial \bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{\partial \theta^*} \frac{\partial \theta^*}{\partial \underline{\theta}} \tag{64}$$

Thus:

$$\frac{d\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{d\underline{\theta}} = -\left(\phi^*(\underline{\theta})(\tau(\underline{\theta}) - \bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})) + \phi^*(\theta^*) \left(\bar{\tau}(\theta^*, \underline{\theta}, \alpha) - \tau(\theta^*)\right) \frac{\partial \theta^*}{\partial \underline{\theta}} \right) \tag{65}$$

According to Corollary **??**, we have $\frac{d\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{d\underline{\theta}} < 0$, and according to Proposition 4, we have $\partial\theta^*/\partial\underline{\theta} < 0$ for $\underline{\theta} < \underline{\theta}^{tr}(c)$ small enough. Hence, using eq.(64), we deduce that for $\underline{\theta}$ close to 0 we have

$$0 < -\frac{\partial \theta^*}{\partial \underline{\theta}} < \phi(\underline{\theta}) \times \frac{\overbrace{\tau(\underline{\theta}) - \bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}^{(*)}}{\phi(\theta^*) \left(\bar{\tau}(\theta^*, \underline{\theta}, \alpha) - \tau(\theta^*)\right)}. \tag{66}$$

The term $(*)$ is dominated by the term $\tau(\underline{\theta})$ for $\underline{\theta}$ small enough. Then, for $\underline{\theta}$ small, there is a constant $K_1 > 0$ such that

$$0 < -\frac{\partial \theta^*}{\partial \underline{\theta}} < K_1 \phi(\underline{\theta}) \tau(\underline{\theta}). \tag{67}$$

and therefore, inserting inequality (67) in equation (64), we obtain that there exists a constant K_2 such that

$$0 < -\frac{d\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{d\underline{\theta}} < K_2 \phi(\underline{\theta}) \tau(\underline{\theta}). \tag{68}$$

Next, we compute the derivative of the second moment in equilibrium and obtain

$$\frac{dm_2(\theta^*, \underline{\theta}, \alpha)}{d\underline{\theta}} = -\left(\phi^*(\underline{\theta}) \left(\tau^2(\underline{\theta}) - m_2(\theta^*, \underline{\theta}, \alpha)\right) + \phi^*(\theta^*) \left(m_2(\theta^*, \underline{\theta}, \alpha) - \tau^2(\theta^*)\right) \frac{\partial \theta^*}{\partial \underline{\theta}} \right) \tag{69}$$

As the order of magnitude of $\partial\theta^*/\partial\underline{\theta}$ is (at best) $\phi(\underline{\theta})\tau(\underline{\theta})$, we deduce from the previous equation that:

$$\frac{dm_2(\theta^*, \underline{\theta}, \alpha)}{d\underline{\theta}} \sim -\phi^*(\underline{\theta})\tau^2(\underline{\theta}), \tag{70}$$

when $\underline{\theta}$ is small. Hence, around $\underline{\theta} = 0$, $\frac{dm_2(\theta^*, \underline{\theta}, \alpha)}{d\underline{\theta}}$ dominates $\frac{d\bar{\tau}(\theta^*, \underline{\theta}, \hat{\alpha})}{d\underline{\theta}}$ by an order of magnitude. Indeed, the ratio between the second derivative and the first is bounded by $1/\tau(\underline{\theta})$, which goes to zero when $\underline{\theta}$ goes to zero.

Proof of Corollary 4. Direct from the arguments in the text.

Proof of Corollary 5. Direct from the arguments in the text.

Proof of Â¨Proposition 6. To be written.