Online Appendix

Equilibrium Data Mining

(not intended for publication)

Jérôme Dugast* Thierry Foucault†

March 17, 2023

*Université Paris Dauphine - PSL. 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

Contents

I. Additional Proofs

A Supplement to the proof of Proposition 5

Remember from eq.(63) in the text that:

$$r(\tau,\tau^*)=\frac{g(\tau,\tau^*)}{g(\tau^*,\tau^*)}=\left(\frac{\chi^2\tau^*+\chi^2+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})}{\chi^2\tau+\chi^2+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})}\right)^{\frac{1}{2}}, \tag{1}$$

with $\chi^2=\rho^2\sigma_\omega^2\sigma_\eta^2$, and $\bar{\tau}(\tau^*;\tau_{dm}^{max})=\mu^*\,\mathsf{E}_\phi\left[\tau\,|\tau^*\leq\tau\leq\tau_{dm}^{max}\right]+(1-\mu^*)\,\mathsf{E}_\gamma\left[\tau\,|\tau^*\leq\tau\right]$, with $\mu^*=\Gamma(\tau^*)$ we deduce that:

$$\begin{aligned}\frac{\partial r(\tau,\tau^*)}{\partial\tau_{dm}^{max}}=&\frac{\partial\,\mathsf{E}_\phi\left[\tau'|\,\tau^*\leq\tau'\leq\tau_{dm}^{max}\right]}{\partial\tau_{dm}^{max}}\Gamma(\tau^*)\,\mathsf{E}_\phi\left[\tau'|\,\tau^*\leq\tau'\leq\tau_{dm}^{max}\right]\\&\times\frac{\chi^2(\tau-\tau^*)}{\{\chi^2\tau+\chi^2+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})\}^{\frac{3}{2}}}\\&\times\frac{1}{\{\chi^2\tau^*+\chi^2+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})\}^{\frac{1}{2}}}.\end{aligned} \tag{2}$$

First, since

$$\mathsf{E}_\phi\left[\tau'|\,\tau^*\leq\tau'\leq\tau_{dm}^{max}\right]=\frac{\int_{\tau^*}^{\tau_{dm}^{max}}\tau'\phi(\tau')d\tau'}{1-\Phi(\tau^*)}=\frac{\int_{\tau^*}^{\tau_{dm}^{max}}\tau'\psi(\tau')d\tau'}{\Psi(\tau_{dm}^{max})-\Psi(\tau^*)}, \tag{3}$$

one can compute

$$\frac{\partial\,\mathsf{E}_\phi\left[\tau'|\,\tau^*\leq\tau'\leq\tau_{dm}^{max}\right]}{\partial\tau_{dm}^{max}}=\frac{\psi(\tau_{dm}^{max})}{\Psi(\tau_{dm}^{max})-\Psi(\tau^*)}\times\left(\tau_{dm}^{max}-\mathsf{E}_\phi\left[\tau'|\,\tau^*\leq\tau'\leq\tau_{dm}^{max}\right]\right). \tag{4}$$

Second, we observe that the term in the second line of eq.(2) can be bounded below as follows:

$$\begin{aligned}&\frac{\chi^2(\tau-\tau^*)}{\{\chi^2\tau+\chi^2+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})\}^{\frac{3}{2}}}\\&=\frac{1}{\tau^{1/2}}\times\frac{\chi^2(1-\tau^*\tau^{-1})}{\{\chi^2+\chi^2\tau^{-1}+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})\tau^{-1}\}^{\frac{3}{2}}}\\&>\frac{1}{(\tau_{dm}^{max})^{1/2}}\times\frac{\chi^2(1-\tau^*\tau^{-1})}{\{\chi^2+\chi^2(\tau^*)^{-1}+\bar{\tau}^2(\tau^*;\tau_{dm}^{max})(\tau^*)^{-1}\}^{\frac{3}{2}}}>0.\end{aligned} \tag{5}$$

Thus:

$$\frac{\int_{\tau^*}^{\tau_{dm}^{max}} \frac{\partial r}{\partial \tau_{dm}^{max}} \psi(\tau) d\tau}{\psi(\tau_{dm}^{max})} > \left(K(\tau^*, \tau_{dm}^{max}) \frac{\tau_{dm}^{max} - \mathsf{E}_\phi \left[\tau' \middle| \tau^* \leq \tau' \leq \tau_{dm}^{max} \right]}{(\tau_{dm}^{max})^{1/2}} \left(1 - \tau^* \, \mathsf{E}_\phi \left[(\tau')^{-1} \middle| \tau^* \leq \tau' \leq \tau_{dm}^{max} \right] \right) \right) \tag{6}$$

where

$$\begin{aligned} K(\tau^*, \tau_{dm}^{max}) = \Gamma(\tau^*) \, \mathsf{E}_\phi \left[\tau' \middle| \tau^* \leq \tau' \leq \tau_{dm}^{max} \right] \\ \times \frac{\chi^2}{\left\{ \chi^2 + \chi^2 (\tau^*)^{-1} + \bar{\tau}^2(\tau^*; \tau_{dm}^{max}) (\tau^*)^{-1} \right\}^{\frac{3}{2}}} \\ \times \frac{1}{\left\{ \chi^2 \tau^* + \chi^2 + \bar{\tau}^2(\tau^*; \tau_{dm}^{max}) \right\}^{\frac{1}{2}}} > 0. \end{aligned} \tag{7}$$

By Assumption A.1, $\mathsf{E}_\phi \left[\tau' \middle| \tau^* \leq \tau' \leq \tau_{dm}^{max} \right]$ exists (does not diverge) for all τ^*. Moreover, $0 < \tau^* < \tau_{dm}^{max}$ for all values of τ_{dm}^{max} (Proposition 2). Thus, $\mathsf{E}_\phi \left[\tau' \middle| \tau^* \leq \tau' \leq \tau_{dm}^{max} \right]$ is bounded, $K(\tau^*, \tau_{dm}^{max})$ admits a lower bound that is strictly positive, and $\tau^* \, \mathsf{E}_\phi \left[(\tau')^{-1} \middle| \tau^* \leq \tau' \leq \tau_{dm}^{max} \right]$ admits an upper bound that is strictly lower than one. We deduce from eq.(6) that $\frac{\int_{\tau^*}^{\tau_{dm}^{max}} \frac{\partial r}{\partial \tau_{dm}^{max}} \psi(\tau) d\tau}{\psi(\tau_{dm}^{max})}$ goes to ∞ when τ_{dm}^{max} goes to ∞.

II. Extensions

A Searching predictors with recall

In the baseline model, we assume that data miners' search for predictors is without recall: When they decide to stop searching, they must trade on the predictor they just discovered. An alternative assumption is that when they stop searching, data miners can use the best of all predictors found until they stop (i.e., they can "recall" predictors found in the past). In this section, we show that this extra flexibility does not affect the equilibrium stopping rule, τ^*.

Step 1. To do so, we first derive the continuation value of a speculator who follows an arbitrary stopping rule $\widehat{\tau}$ when other data miners follow the stopping rule τ^*. A key difference with the baseline case is that this continuation value in a given exploration round depends on the best predictor she found until this round. Let denote this best predictor by τ^{best}. Observe that it must be the case that $\tau^{best} \leq \widehat{\tau}$ (if the inequality

was not satisfied the speculator would have stopped searching in a previous round; a contradiction). The expected utility of launching a new exploration for the speculator is:

$$J(\tau^{best}, \widehat{\tau}, \tau^*) = \exp(\rho c)\left[\left(\int_{\widehat{\tau}}^{\tau_{dm}^{max}} g(\tau,\tau^*)\phi(\tau)d\tau + \int_{\tau^{best}}^{\widehat{\tau}} J(\tau,\widehat{\tau},\tau^*)\phi(\tau)d\tau\right) + \left(1 - \int_{\tau^{best}}^{\tau_{dm}^{max}} \phi(\tau)d\tau\right) J(\tau^{best},\widehat{\tau},\tau^*)\right] \tag{8}$$

Differentiating both sides of eq.(8) with respect to τ^{best} implies that:

$$\begin{aligned} &\frac{\partial J}{\partial \tau^{best}}\left(1 - \exp(\rho c)\left(1 - \int_{\tau^{best}}^{\tau_{dm}^{max}} \phi(\tau)d\tau\right)\right) \\ &= \exp(\rho c)\left[-J(\tau^{best},\widehat{\tau},\tau^*)\phi(\tau^{best}) + \phi(\tau^{best})J(\tau^{best},\widehat{\tau},\tau^*)\right] = 0 \end{aligned} \tag{9}$$

Thus, $\frac{\partial J(\tau^{best},\widehat{\tau},\tau^*)}{\partial \tau^{best}} = 0$. Hence, $J(\tau^{best},\widehat{\tau},\tau^*)$ does not depend on τ^{best} for $\tau^{best} < \widehat{\tau}$. This means that for $\tau^{best} < \tau < \widehat{\tau}$, $J(\tau,\widehat{\tau},\tau^*) = J(\tau^{best},\widehat{\tau},\tau^*)$. Thus, eq.(8) implies:

$$J(\tau^{best},\widehat{\tau},\tau^*) = J(\widehat{\tau},\tau^*) = \exp(\rho c)\left[\int_{\widehat{\tau}}^{\tau_{dm}^{max}} g(\tau,\tau^*)\phi(\tau)d\tau + \left(1 - \int_{\widehat{\tau}}^{\tau_{dm}^{max}} \phi(\tau)d\tau\right) J(\widehat{\tau},\tau^*)\right], \tag{10}$$

Hence, the continuation value of the speculator for any $\tau^{best} \leq \widehat{\tau}$ is the same as in the baseline version of the model. Intuitively, the reason is that there is no limit to the number of explorations that a speculator can perform. Hence, the possibility for a speculator to "store" predictors has no value since the speculator keeps searching as long as the quality of the predictors she found is above $\widehat{\tau}$.

Step 2. Now, suppose that there exists a solution, $\widehat{\tau}$, to the following the indifference condition

$$g(\widehat{\tau},\tau^*) = J(\widehat{\tau},\widehat{\tau},\tau^*). \tag{11}$$

We show below that $\widehat{\tau}$ is the optimal stopping rule of the speculator when other data miners' stopping rule τ^*. According to the one-shot deviation principle, a necessary and sufficient condition for $\widehat{\tau}$ to be optimal is that a one-shot deviation from this policy is not optimal. First, we show that it is always optimal to keep searching when $\tau < \widehat{\tau}$ (thus a

one shot deviation that consists in stopping is not optimal), because

$$g(\tau, \tau^*) < g(\widehat{\tau}, \tau^*) = J(\widehat{\tau}, \widehat{\tau}, \tau^*) = J(\tau, \widehat{\tau}, \tau^*), \tag{12}$$

where the first inequality follows from the fact that $g(\tau, \tau^*)$ increases with τ, the second equality follows from eq.(11), and the last equality follows from the fact that $J(\tau, \widehat{\tau}, \tau^*)$ does not depend on τ for $\tau \leq \widehat{\tau}$ (see Step 1).

Second, consider a case where the speculator's obtains a predictor of type $\tau \in [\widehat{\tau}, \tau_{dm}^{max}]$. If stopping when $\tau > \widehat{\tau}$ is the optimal policy then a one shot deviation from this policy is not optimal. We show here that this is indeed the case. The expected continuation value of the speculator with such a one shot deviation is:

$$X(\tau, \tau^*) = \exp(\rho c) \left[\int_{\tau}^{\tau_{dm}^{max}} g(\tau', \tau^*) \phi(\tau') d\tau' + \left(1 - \int_{\tau}^{\tau_{dm}^{max}} \phi(\tau') d\tau' \right) g(\tau, \tau^*) \right]. \tag{13}$$

Indeed, either the speculator finds an even better predictor in the next exploration round (first term in the R.H.S of eq.(13)) or she does not and then her policy will command to stop at the next round with the best predictor found so far, i.e., the predictor of quality τ (second term in eq.(13)). Notice that $X(\widehat{\tau}, \tau^*) = J(\widehat{\tau}, \widehat{\tau}, \tau^*) = g(\widehat{\tau}, \tau^*)$ which shows that a speculator with $\tau = \widehat{\tau}$ is indifferent between deviating or not. Now, we show that:

$$\forall\, \tau > \widehat{\tau},\ g(\tau, \tau^*) > X(\tau, \tau^*). \tag{14}$$

As $g(\tau, \tau^*) < 0$, using the expression for $X(\tau, \tau^*)$, this inequality is equivalent to

$$\forall\, \tau > \widehat{\tau},\ \exp(-\rho c) < Y(\tau, \tau^*), \tag{15}$$

where $Y(\tau, \tau^*) = \int_{\tau}^{\tau_{dm}^{max}} \frac{g(\tau', \tau^*)}{g(\tau, \tau^*)} \phi(\tau') d\tau' + 1 - \int_{\tau}^{\tau_{dm}^{max}} \phi(\tau') d\tau'$. Now:

$$\frac{\partial Y}{\partial \tau} = \int_{\tau}^{\tau_{dm}^{max}} \frac{\partial}{\partial \tau} \left(\frac{g(\tau', \tau^*)}{g(\tau, \tau^*)} \right) \phi(\tau') d\tau'. \tag{16}$$

The expected utility $g(\tau, \tau^*)$ is negative, and is increasing in τ. Therefore $\frac{g(\tau', \tau^*)}{g(\tau, \tau^*)}$ is positive and also increasing in τ. Thus, $\frac{\partial Y}{\partial \tau} > 0$ and therefore:

$$\forall\, \tau > \widehat{\tau},\ \exp(-\rho c) = Y(\widehat{\tau}, \tau^*) < Y(\tau, \tau^*), \tag{17}$$

or equivalently $\forall\, \tau > \widehat{\tau},\ g(\tau, \tau^*) > X(\tau, \tau^*)$. Thus, $\widehat{\tau}$ is the optimal stopping rule for the speculator. Moreover, in a symmetric equilibrium, it must be the case that $\widehat{\tau}(\tau^*) = \tau^*$. This means that:

$$g(\tau^*, \tau^*) = J(\tau^*, \tau^*, \tau^*). \tag{18}$$

As the continuation value $J(\tau^*, \tau^*, \tau^*)$ is identical to that in the baseline model (see Step 1), the stopping rule τ^* is the same. Thus, the results are unchanged.

B Uniqueness of Equilibrium with Markovian Search Strategies

In our model, before the $(t+1)^{th}$ search stage , an asset manager strategy depends, in full generality, on the number of completed search round t, and on the search history $\mathcal{H}_t = \{0, \tau_1, \tau_2, \ldots, \tau_t\}$. Define $\mathcal{M}_t \subset \mathcal{H}_t$, the set of all drawn predictors that can still be used (e.g., without recall, $\mathcal{M}_t = \tau_t$). Prior to round $t+1$, if an asset manager decides to stop searching, she will use her best possible predictor $\tau_t^{best} = \sup \mathcal{M}_t$. Hence, an asset manager strategy is a mapping, S,

$$S : (\mathcal{H}_t, \mathcal{M}_t) \mapsto d_{t+1} \in \{\text{stop}, \text{search}\} \tag{19}$$

In this paper focus on *Markovian* strategies, where the strategy does only depend on τ_t^{best}, or simply τ^{best}, which is the only pay-off relevant state variable, that is

$$S : \tau^{best} \mapsto d \in \{\text{stop}, \text{search}\} \tag{20}$$

Therefore, the search strategy is equivalently a set $\Omega \subset [0, \tau_{dm}^{max}]$ such that $S(\tau^{best}) = \text{stop}$ if and only if $\tau^{best} \in \Omega$. Ω shall have a non-zero measure, and be compact. Therefore it is a segment, or a reunion of segments, of $[0, \tau_{dm}^{max}]$.

Given the search strategies of all asset managers, one can always define the average predictors quality, $\bar{\tau}$, in the trading stage, and the expected utility of trading using a predictor of quality τ:

$$g(\tau) = -\left(1 + \frac{\tau}{1 + \frac{\bar{\tau}^2}{\rho^2 \sigma_\omega^2 \sigma_\nu^2}}\right)^{-\frac{1}{2}} \tag{21}$$

Without recall, when an asset manager goes for another round of search, she obtains the continuation value, J, determined as follows

$$J = \exp(\rho c)\left[\int_{\Omega} g(\tau')\phi(\tau')d\tau' + \left(1 - \int_{\Omega}\phi(\tau')d\tau'\right)J\right]. \tag{22}$$

For any $\tau \in [0, \tau_{dm}^{max}] \setminus \Omega$,we have $J > g(\tau)$. And, for any $\tau \in \Omega$, we have $J \leq g(\tau)$. Since $g(\tau)$ is an increasing and continuous function of τ, there exists a $\widehat{\tau}$ such that $J = g(\widehat{\tau})$. It then implies that $J \geq g(\tau)$ if and only if $\tau \leq \widehat{\tau}$ and thus $\Omega = [\widehat{\tau}, \tau_{dm}^{max}]$.

When there is recall, all past draws may in principle still be usable, that is $\mathcal{M}_t = \mathcal{H}_t$ and $\tau_t^{best} = \sup \mathcal{H}_t$. We want to show that $\Omega = [\widehat{\tau}, \tau_{dm}^{max}]$, that is has no "hole". Assume that this is not the case and that one can find $\tau_1 < \tau_2$ such that $\Omega \cap [0, \tau_1] \neq \emptyset$, $\Omega \cap [\tau_2, \tau_{dm}^{max}] \neq \emptyset$, and $\Omega \cap (\tau_1, \tau_2) = \emptyset$. One can show that the continuation value of going for another search round, $J(\tau^{best})$, is constant on (τ_1, τ_2). Let $\tau^{best} = \tau \in (\tau_1, \tau_2)$ and compute

$$\begin{aligned} J(\tau) =& \exp(\rho c)\left[\Phi(\tau)J(\tau) + \int_{[\tau_2, \tau_{dm}^{max}]\cap\Omega} g(\tau')\phi(\tau')d\tau'\right] \\ &+ \exp(\rho c)\left[\int_{\tau}^{\tau_2} J(\tau')\phi(\tau')d\tau' + \int_{[\tau_2, \tau_{dm}^{max}]\setminus\Omega} J(\tau')\phi(\tau')d\tau'\right] \end{aligned} \tag{23}$$

Taking the derivative with respect to τ, one obtains

$$\frac{\partial J}{\partial \tau}\left(1 - \exp(\rho c)\Phi(\tau)\right) = \exp(\rho c)\left[\phi(\tau)J(\tau) - J(\tau)\phi(\tau)\right] = 0. \tag{24}$$

In addition, J must be weakly increasing in τ because having a larger τ^{best} shall not reduce the expected utility of the data miner. We must also have $J(\tau) > g(\tau)$ on (τ_1, τ_2), and $J(\tau) \leq g(\tau)$ on $\Omega \cap [0, \tau_1]$ and on $\Omega \cap [\tau_2, \tau_{dm}^{max}]$.

Now consider the alternative strategy $\tilde{S}$ associated to the set $\tilde{\Omega} = [\tau_1, \tau_{dm}^{max}] \cap \Omega$. In that case, the continuation value $\tilde{J}$ is equal to J on $[\tau_2, \tau_{dm}^{max}]$ and $\tilde{J}$ is constant on $[0, \tau_2]$. Moreover, $\tilde{J}$ has the exact same definition than, and thus is equal to, J on (τ_1, τ_2). As J is weakly increasing, it implies that $\tilde{J}(\tau) \geq J(\tau)$ on $[0, \tau_2]$. However, for some $\tau_0 \in \Omega \cap [0, \tau_1]$, we must have $J(\tau_0) < g(\tau_0) \leq \tilde{J}(\tau_0)$. Hence, for $\tau \in [0, \tau_0]$, we have $\tilde{J}(\tau) > J(\tau)$. The strategy $\tilde{S}$ yields a larger expected utility than S. It contradicts the optimality of Ω. Thus, Ω must be a single segment within $[0, \tau_{dm}^{max}]$.

Moreover, $\tau_{dm}^{max} \in \Omega$. And there must be τ_H close enough to τ_{dm}^{max}, such that $[\tau_H, \tau_{dm}^{max}] \subset \Omega$ because the probability of finding a better predictor (and the expected utility gain) than $\tau^{best} \in [\tau_H, \tau_{dm}^{max}]$ is so low that it cannot out weight the search cost. Indeed, one can compute the expected utility of going for a single search round as

$$\exp(\rho c)\left[\Phi(\tau^{best})g(\tau^{best}) + \int_{\tau^{best}}^{\tau_{dm}^{max}} g(\tau')\phi(\tau')d\tau'\right], \tag{25}$$

which is lower than $g(\tau^{best})$ for τ^{best} high enough. Thus, $\Omega = [\widehat{\tau}, \tau_{dm}^{max}]$ for some $\widehat{\tau}$.

C Searching predictors by combining signals

In this section, we formalize more explicitly the process by which data miners obtain their predictors. Our approach makes clear why the quality of a speculator's predictor in a given round is not necessarily greater than in a previous round, as assumed in our baseline model.

Assume that in a given round data miners can use N variables (signals) s_j to predict the asset payoff ω. We formalize each variable as being a signal about the asset payoff. Specifically, we assume that

$$s_j = \omega + (\tau_j)^{-\frac{1}{2}}\varepsilon_j, \quad j \in \{1, ..., N\} \tag{26}$$

where the ε_j's have a normal distribution with mean zero and precision $\theta_\omega = 1/\sigma_\omega^2$. Importantly, we assume that τ_j is specific to each variable used in a given round and can therefore vary across variables. Moreover, we assume that τ_j is drawn from some distribution. The key assumption is that the number of variables used to predict the asset payoff in a given exploration round is fixed (N can be large but it cannot increase as new rounds are launched). This means that in a given round, a speculator must replace at least one of the variable used in the past by a new variable whose predictive quality (τ_j) is unknown (it is discovered in the exploration round). In reality, research teams may fix the number of variables used in their predictive models to avoid overfitting. Thus, we think that the assumption that N is fixed (it can be large) is reasonable. As explained below, this assumption implies that the quality of a predictor in a given round is not necessarily larger as in previous rounds (as what would happen if one could retain

variables from past explorations and add new ones). Moreover, we show below that the quality τ of a predictor in a given round is the sum of the quality of each variable used to form the predictor (i.e., $\tau = \sum_{j=1}^{j=N} \tau_j$).

To see this formally, fix the variables used in the n^{th} round of exploration and denote by $\tau_j(n)$ the realization of the precision for signal s_j in round n. Similarly, let $\tau(n) = \sum_{j=1}^{j=N} \tau_j(n)$ be the sum of these precisions in the n^{th} round. Using standard properties of normally distributed variables, we deduce that:

$$\mathsf{E}(\omega \mid s_1(n), s_2(n), ..., s_N(n)) = \sum_{j=1}^{j=N} \mu_j(n) s_j, \tag{27}$$

with $\mu_j(n) = \frac{\tau_j(n)}{1+\tau(n)}$. As all variables are normally distributed, $\mu_j(n)$ is the coefficient that would be obtained for the j^{th} variable in running a regressing of ω on the N variables used in round n. The predictor obtained in this round is just the predicted value of this regression, i.e., $\mathsf{E}(\omega \mid s_1(n), s_2(n), ..., s_N(n))$. Thus, the predictor obtained in round n is:

$$s_{\tau(n)} = \sum_{j=1}^{j=N} \mu_j(n) s_j = \left(\sum_j \mu_j(n) \right) \omega + \sum_j \mu_j(n) (\tau_j(n))^{-\frac{1}{2}} \varepsilon_j. \tag{28}$$

Alternatively, one can use as predictor (its informativeness is identical):

$$\hat{s}_{\tau(n)} = \omega + \sum_j \frac{\mu_j(n)}{\sum_j \mu_j(n)} (\tau_j(n))^{-\frac{1}{2}} \varepsilon_j. \tag{29}$$

Using the definition of $\mu_j(n)$, this can be rewritten:

$$\hat{s}_{\tau(n)} = \omega + (\tau(n))^{-\frac{1}{2}} \varepsilon_{\tau(n)}. \tag{30}$$

where $\varepsilon_{\tau(n)} = \sum_j \frac{(\tau_j(n))^{\frac{1}{2}}}{\left(\sum_j \tau_j(n)\right)^{\frac{1}{2}}} \varepsilon_j$. Thus, this process generates predictors that have exactly the structure of those considered in our model (Section II.A in this appendix shows that this specification of predictors is equivalent to that considered in the baseline model). Thus, $\tau(n)$ is the quality of the predictor built with N signals in the n^{th} round. Note that the theoretical R^2 of the regression of ω on the N variables used in round n (i.e., $1 - \mathsf{Var}[\omega \mid \hat{s}_{\tau(n)}] / \mathsf{Var}[\omega]$) is equal to $\tau(n)(1+\tau(n))^{-1}$. Thus, the higher the quality of a predictor, the higher the R^2 of a regression of the asset payoff on the predictor.

Now suppose that a speculator decides to launch a $(n+1)^{th}$ round of exploration by considering a different set of variables. As at least one variable is different, $\tau(n+1)$ for the predictor built with these variables will be different from $\tau(n)$. Moreover as the precision of the signals used in one round are random, the precision of the predictor obtained in the $(n+1)^{th}$ round can be larger or smaller than that of the previous predictor. For instance, suppose that in each round of exploration, τ_j follows a Gamma distribution with parameters γ_0 and γ_1. Then, if the speculator changes all the variables used in the $(n+1)^{th}$ round of exploration, the probability that $\tau(n+1)$ is smaller than $\tau(n)$ is $G(\tau(n); N\gamma_0, \gamma_1)$ where $G(.)$ is the cumulative probability distribution of a Gamma distribution with parameters $N\gamma_0$ and γ_1. Clearly, the same possibility arises even if the speculator only considers changing a subset of all variables used in the previous round.

D No risk neutral dealers

In this section, we show how one can extend our baseline model to a setting without risk neutral market makers (as in Grossman and Stiglitz (1980) for instance). As explained below, in this case, the equilibrium is unchanged. In particular, data miners' equilibrium search intensity is unchanged and investors' capital allocation decisions are as in the baseline model. It follows that the implications of the baseline model still hold in this case. The main difference with the baseline model is that the expected risk premium of the asset is not zero, so that managers' expected profit is non zero if they have no private information.

Let $S \geq 0$ be the supply of the asset. The equilibrium of the model is defined as in Section III.D, except that now the equilibrium asset price, p^*, must satisfy the following market clearing condition:

$$\int x^*(s_\tau, p^*)di + \eta = S. \tag{31}$$

D.1 Equilibrium.

We first describe all equilibrium variables in our model without risk neutral dealers (these variables are defined as in the model; we do not repeat the definition here) and we then

explain how we obtain them in the next subsection. The equilibrium price of the asset is:

$$p^* = \frac{\bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}{1 + \bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}\xi - \frac{\rho\sigma_\omega^2}{1 + \bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}S, \text{ with } \xi = \omega + \rho\sigma_\omega^2\bar{\tau}^{-1}\eta. \tag{32}$$

The informativeness of the equilibrium price and the expected risk premium on the asset are respectively

$$\mathcal{I}(\tau^*; \tau_{dm}^{max}) = \mathsf{Var}[\omega|p^*]^{-1} = \frac{1}{\sigma_\omega^2}\left(1 + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}\right), \text{ and } \mathsf{E}[\omega - p^*] = \frac{\rho\sigma_\omega^2}{1 + \bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}S. \tag{33}$$

The optimal demand for the risky asset of an asset manager with signal s_τ is:

$$x^*(s_\tau, p^*) = \frac{\mathsf{E}[\omega|s_\tau, p^*] - p^*}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]} = \frac{\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]} + \frac{\mathsf{E}[\omega|p^*] - p^*}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}, \tag{34}$$

with

$$\frac{\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]} = \frac{\tau}{\rho\sigma_\omega^2}\left(s_\tau - \mathsf{E}[\omega|p^*]\right). \tag{35}$$

Thus, an asset manager's demand has two components. One that exploits her private information (the first component) and one that exploits the fact that the asset expected risk premium is not zero (the second component). The second component is zero in the presence of risk neutral dealers.

The expected utility from trading with a signal of quality τ is:

$$g(\tau, \tau^*) = \left(1 + \frac{\tau}{1 + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}\right)^{-1/2} \times g_U(\tau^*) \tag{36}$$

where $g_U(\tau^*)$ is the expected utility of trading without an informative signal, that is

$$g_U(\tau^*) = -\,\mathsf{E}\left[\exp\left(-\frac{(\mathsf{E}[\omega|p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|p^*]}\right)\right]. \tag{37}$$

In our baseline model, $g_U(\tau^*) = 1$ because the expected risk premium of the asset is zero ($p^* = \mathsf{E}[\omega|p^*]$) due to dealers' risk neutrality.

The expected utility of searching for a predictor with a stopping rule $\widehat{\tau}$, $J(\widehat{\tau}, \tau^*)$, has therefore the same definition as in the paper, up to the factor $-g_U(\tau^*)$, and the equilibrium condition $g(\tau^*, \tau^*) = J(\tau^*, \tau^*)$ yields therefore exactly the same equilibrium

stopping rule for data miners, τ^*.

The expected profit obtained by an asset manager with a signal of precision τ is:

$$\bar{\Pi}(\tau) = \mathsf{E}[x^*(s_\tau, p^*)(\omega - p^*)|\tau] = \underbrace{\mathsf{E}\left[\frac{(\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*])(\omega - \mathsf{E}[\omega|p^*])}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}\middle|\tau\right]}_{VA(\tau)} + \underbrace{\mathsf{E}\left[\frac{(\mathsf{E}[\omega|p^*] - p^*)^2}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}\middle|\tau\right]}_{RT(\tau)}. \tag{38}$$

where

$$VA(\tau) = \frac{\tau}{\rho\sigma_\omega^2 \mathcal{I}(\tau^*; \tau_{dm}^{max})}. \tag{39}$$

Thus, the expected profit of an asset manager has two components. The first component reflects the expected profit ($VA(\tau)$) that she derives from her private information, s. The expression for this component is identical to that obtain for asset managers' expected profit in Section VI (see eq.(27) in the paper). It is specific to each asset manager since it depends on the precision, τ, of her signal. The second component ($RT(\tau)$) reflects the expected profit that an asset manager can earn even if she is uninformed, by collecting the asset risk premium. As the expression for the value added component is identical to that in the paper, all our results in Section VI also obtains when one measures the performance of asset managers by $VA(\tau)$. Empirically, one can measure $VA(\tau)$ by adjusting an asset managers' expected profit by the asset expected excess return times the asset managers' holding of the risky asset x^* ($x^*(\mathsf{E}[\omega|p^*] - p^*) = x^* p^* \times \mathsf{E}[r_e|p^*]$ where r_e is the excess return on the risky asset; remember that the risk free return is normalized to zero).

D.2 Proofs.

First, we conjecture and verify later that the equilibrium price is linear in ω, η and S, that is

$$p^* = a\omega + b\eta - dS = a\xi - dS, \text{ with } \xi \equiv \omega + \frac{b}{a}\eta. \tag{40}$$

Thus, p and ξ are observationally equivalent. Let $\theta_\omega \equiv 1/\sigma_\omega^2$ be the precision of asset managers' prior about the asset pay-off ω, and $\theta_\xi \equiv a^2/(b^2\sigma_\eta^2)$ is the precision of ξ as a signal about ω. Using standard calculations in the CARA gaussian framework, we obtain that the optimal demand for the risky asset of an asset manager with signal s_τ is:

$$x^*(s_\tau, p^*) = \frac{\mathsf{E}[\omega|s_\tau, p^*] - p^*}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}, \tag{41}$$

As all variables are normally distributed and ε_i and η (the noises in s_τ and ξ) are independent, standard calculations (using eq.(40) and the fact that the precision of s_τ is $\tau\theta_\omega$) yield:

$$\mathsf{E}[\omega|s_\tau, p^*] = \frac{\tau\theta_\omega s_\tau + \theta_\xi \xi}{\theta_\omega + \tau\theta_\omega + \theta_\xi}. \tag{42}$$

and

$$\mathsf{Var}[\omega|s_\tau, p^*] = \frac{1}{\theta_\omega + \tau\theta_\omega + \theta_\xi}. \tag{43}$$

Therefore an asset manager's demand is

$$x^*(s_\tau, p^*) = \frac{1}{\rho}\left(\tau\theta_\omega s_\tau + \theta_\xi \xi - (\theta_\omega + \tau\theta_\omega + \theta_\xi)p^*\right), \tag{44}$$

and the market clearing conditions follows as

$$\frac{1}{\rho}\left(\bar{\tau}\theta_\omega \omega + \theta_\xi \xi - (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)p^*\right) + \eta = S. \tag{45}$$

The latter can be rewritten as

$$(\bar{\tau}\theta_\omega + \theta_\xi)\omega + \left(\rho + \theta_\xi \frac{b}{a}\right)\eta - \rho S = (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)p^*. \tag{46}$$

Thus, our conjecture regarding the equilibrium price is satisfied if a, b and d are such that:

$$\begin{aligned} (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)a &= \bar{\tau}\theta_\omega + \theta_\xi \\ (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)b &= \rho + \theta_\xi \frac{b}{a} \\ (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)d &= \rho \end{aligned} \tag{47}$$

Thus,

$$\begin{aligned} (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)b &= \rho + \frac{\theta_\xi}{\bar{\tau}\theta_\omega + \theta_\xi}(\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)b \\ \Rightarrow (\theta_\omega + \bar{\tau}\theta_\omega + \theta_\xi)b &= \frac{\rho(\bar{\tau}\theta_\omega + \theta_\xi)}{\bar{\tau}\theta_\omega}. \end{aligned} \tag{48}$$

This implies that

$$\xi = \omega + \frac{b}{a}\eta = \omega + \rho\sigma_\omega^2 \bar{\tau}^{-1}\eta, \text{ and, } \theta_\xi = \frac{a^2}{b^2\sigma_\eta^2} = \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^4\sigma_\eta^2}, \tag{49}$$

exactly as in Proposition 1 in the paper, and finally that

$$\begin{aligned} a &= \frac{\bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}{1 + \bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}} \\ \text{and } d &= \frac{\rho\sigma_\omega^2}{1 + \bar{\tau} + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}} \end{aligned} \tag{50}$$

Calculations yield

$$\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*] = \frac{\tau\theta_\omega s_\tau + \theta_\xi\xi}{\theta_\omega + \tau\theta_\omega + \theta_\xi} - \frac{\theta_\xi\xi}{\theta_\omega + \theta_\xi} = \frac{\tau\theta_\omega\left(s_\tau - \frac{\theta_\xi}{\theta_\omega+\theta_\xi}\xi\right)}{\theta_\omega + \tau\theta_\omega + \theta_\xi} = \frac{\tau\theta_\omega\left(s_\tau - \mathsf{E}[\omega|p^*]\right)}{\theta_\omega + \tau\theta_\omega + \theta_\xi}, \tag{51}$$

and therefore

$$x^*(s_\tau, p^*) = \frac{\tau}{\rho\sigma_\omega^2}\left(s_\tau - \mathsf{E}[\omega|p^*]\right) + \frac{\mathsf{E}[\omega|p^*] - p^*}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}, \tag{52}$$

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

$$\begin{aligned} &\mathsf{E}[-\exp(-\rho x^*(s_\tau, p^*)(\omega - p^*)) \mid s_\tau, p^*] \\ &= -\,\mathsf{E}\left[\exp\left(-\rho\left(x^*(s_\tau, p^*)(\mathsf{E}[\omega \mid s_\tau, p^*] - p^*) - \frac{\rho(x^*(s_\tau, p^*))^2}{2}\,\mathsf{Var}[\omega \mid s_\tau, p^*]\right)\right)\right]. \end{aligned} \tag{53}$$

Substituting $x^*(s_\tau, p^*)$ by its expression, we deduce that:

$$\mathsf{E}\left[-\exp(-\rho x^*(s_\tau, p^*)(\omega - p^*)) \mid s_\tau, p^*\right] = -\exp\left(-\frac{(\mathsf{E}[\omega|s_\tau, p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|s_\tau, p^*]}\right) \tag{54}$$

Thus:

$$\begin{aligned} g(\tau, \tau^*) &= -\,\mathsf{E}\left[\exp\left(-\frac{(\mathsf{E}[\omega|s_\tau, p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|s_\tau, p^*]}\right)\right] \\ &= -\,\mathsf{E}\left[\mathsf{E}\left[\exp\left(-\frac{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - p^*|p^*\right]}{2\,\mathsf{Var}[\omega|s_\tau, p^*]}\left(\frac{\mathsf{E}[\omega|s_\tau, p^*] - p^*}{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - p^*|p^*\right]^{1/2}}\right)^2\right)\middle| p^*\right]\right] \end{aligned} \tag{55}$$

Conditional on p^*, the variable $\mathsf{E}[\omega|s_\tau, p^*] - p^*$ has a mean equal to $\mathsf{E}[\omega|p^*] - p^*$, and a variance equal to $\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - p^*|p^*\right] = \mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]\right]$. Next, define

$$Z = \frac{\mathsf{E}[\omega|s_\tau, p^*] - p^*}{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - p^*|p^*\right]^{1/2}} = \frac{\mathsf{E}[\omega|s_\tau, p^*] - p^*}{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]\right]^{1/2}} \tag{56}$$

Conditional on p^*, the variable $Z|p^*$ has a variance equal to 1, and a mean equal to

$$\mathsf{E}[Z|p^*] = \frac{\mathsf{E}[\omega|p^*] - p^*}{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]\right]^{1/2}}. \tag{57}$$

For a normally distributed variable Z with mean m and variance 1, we have (see proof of Theorem 2, eq (A21), in Grossman-Stiglitz (1980))

$$\mathsf{E}[\exp(-tZ^2)] = (1+2t)^{-1/2} \times \exp\left(-\frac{t}{1+2t}m^2\right). \tag{58}$$

Thus,

$$g(\tau, \tau^*) = -\,\mathsf{E}\left[\left(1 + \frac{\mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]\right]}{\mathsf{Var}[\omega|s_\tau, p^*]}\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\tau,p^*]-\mathsf{E}[\omega|p^*]]}{\mathsf{Var}[\omega|s_\tau,p^*]}}{1+\frac{\mathsf{Var}[\mathsf{E}[\omega|s_\tau,p^*]-\mathsf{E}[\omega|p^*]]}{\mathsf{Var}[\omega|s_\tau,p^*]}}\,\mathsf{E}[Z|p^*]^2\right)\right] \tag{59}$$

Using the conditional variance decomposition ($\mathsf{Var}[Y] = \mathsf{E}[\mathsf{Var}[Y|X]|X] + \mathsf{Var}[\mathsf{E}[Y|X]|X]$), we also obtain that

$$\begin{aligned} \mathsf{Var}[\omega|p^*] &= \mathsf{E}[\mathsf{Var}[\omega|s_\tau, p^*]|p^*] + \mathsf{Var}[\mathsf{E}[\omega|s_\tau, p^*]|p^*] \\ &= \mathsf{Var}[\omega|s_\tau, p^*] + \mathsf{Var}\left[\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]\right], \end{aligned} \tag{60}$$

and then

$$g(\tau, \tau^*) = -\left(\frac{\mathsf{Var}[\omega|p^*]}{\mathsf{Var}[\omega|s_\tau, p^*]}\right)^{-1/2} \mathsf{E}\left[\exp\left(-\frac{(\mathsf{E}[\omega|p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|p^*]}\right)\right] \tag{61}$$

Notice that the expected utility of trading without an informative signal is

$$g_U(\tau^*) = -\,\mathsf{E}\left[\exp\left(-\frac{(\mathsf{E}[\omega|p^*] - p^*)^2}{2\,\mathsf{Var}[\omega|p^*]}\right)\right] \tag{62}$$

Moreover, given that

$$\mathsf{Var}[\omega|s_\tau, p^*] = \frac{1}{\theta_\omega + \tau\theta_\omega + \theta_\xi} = \frac{\sigma_\omega^2}{1 + \tau + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}, \text{ and } \mathsf{Var}[\omega|p^*] = \frac{\sigma_\omega^2}{1 + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}} = \mathcal{I}(\tau^*; \tau_{dm}^{max})^{-1}, \tag{63}$$

we finally obtain

$$g(\tau, \tau^*) = \left(1 + \frac{\tau}{1 + \frac{\bar{\tau}^2}{\rho^2\sigma_\omega^2\sigma_\eta^2}}\right)^{-1/2} \times g_U(\tau^*) \tag{64}$$

The expected profit of an asset manager with quality τ is defined as

$$\bar{\Pi}(\tau) = \mathsf{E}[x^*(s_\tau, p^*)(\omega - p^*)|\tau] = \mathsf{E}\left[\frac{\mathsf{E}[\omega|s_\tau, p^*] - p^*}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}(\omega - p^*)\middle|\tau\right] \tag{65}$$

Notice that $\omega - p^*$ can be decomposed into two independent components: $\omega - \mathsf{E}[\omega|p^*]$, and $\mathsf{E}[\omega|p^*] - p^*$. Similarly, $\mathsf{E}[\omega|s_\tau, p^*] - p^*$ can be decomposed into $\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]$, and $\mathsf{E}[\omega|p^*] - p^*$. Therefore the expected profit can be decomposed a value added component, $VA(\tau)$, and a pure risk taking component $RT(\tau)$, as follows

$$\bar{\Pi}(\tau) = \underbrace{\mathsf{E}\left[\frac{\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}(\omega - \mathsf{E}[\omega|p^*])\middle|\tau\right]}_{VA(\tau)} + \underbrace{\mathsf{E}\left[\frac{(\mathsf{E}[\omega|p^*] - p^*)^2}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}\middle|\tau\right]}_{RT(\tau)}. \tag{66}$$

with

$$\begin{aligned}\mathsf{E}\left[\frac{\mathsf{E}[\omega|s_\tau, p^*] - \mathsf{E}[\omega|p^*]}{\rho\,\mathsf{Var}[\omega|s_\tau, p^*]}(\omega - \mathsf{E}[\omega|p^*])\middle|\tau\right] &= \mathsf{E}\left[\frac{\tau\theta_\omega}{\rho}(\underbrace{\omega + \tau^{-1/2}\varepsilon_i}_{=s_\tau} - \mathsf{E}[\omega|p^*])(\omega - \mathsf{E}[\omega|p^*])\right] \\ &= \mathsf{E}\left[\frac{\tau\theta_\omega}{\rho}(\omega - \mathsf{E}[\omega|p^*])^2\right] = \frac{\tau}{\rho\sigma_\omega^2\mathcal{I}(\tau^*; \tau_{dm}^{max})}.\end{aligned} \tag{67}$$

III. A specific distribution for predictors' types

In this section, we present (in (Section III.B) a family of probability distributions for $\psi(.)$ for which the function $F(.)$ can be computed in closed-form . As $F(.)$ can be computed in closed-form, the equilibrium stopping rule (τ^*) as well as various variables of interest that depends on the equilibrium stopping rule (e.g., the mean and the variance of the distribution of the quality of predictors across data miners) can be computed as well (at least numerically). In Section III.C, we focus on the two special cases of this family of distributions that we use in the numerical examples considered in the paper. We first start (in Section III.A) the analysis with a change in variables for predictors' quality, τ. This re-parametrization of the model proves convenient for the rest of the analysis in this section.

A A Parametrization of predictors' quality

For any $\tau \in [0, +\infty)$, there is a unique *predictor's type*, $\theta \in [0, \pi/2]$, such that

$$\tau = \cot^2(\theta) = \frac{\cos^2(\theta)}{\sin^2(\theta)} = \frac{1}{\sin^2(\theta)} - 1 \;\Leftrightarrow\; \theta = \arcsin[(1+\tau)^{-1/2}] \tag{68}$$

Under this parametrization, $\tau(\theta)$ decreases with θ, with $\tau(0) = \infty$ and $\tau(\pi/2) = 0$.

If θ is distributed over $[0, \pi/2]$ according to the distribution $h(.)$ (cdf, $H(.)$), then we obtain the corresponding distribution for τ:

$$\Psi(\tau) = 1 - H\left(\arcsin[(1+\tau)^{-1/2}]\right). \tag{69}$$

Moreover, as $\frac{\partial \arcsin(x)}{\partial x} = (1-x)^{-1/2}$, we deduce that:

$$\psi(\tau) = \frac{1}{2(1+\tau)\tau^{1/2}} h(\arcsin[(1+\tau)^{-1/2})) \tag{70}$$

Then for any $\tau_a < \tau_b$, any function $f(.)$, and by defining $\theta_a = \arcsin[(1+\tau_a)^{-1/2}] > \theta_b = \arcsin[(1+\tau_b)^{-1/2}]$, one has

$$\int_{\tau_a}^{\tau_b} f(\tau)\psi(\tau)d\tau = \int_{\theta_b}^{\theta_a} f(\cot^2(\theta))h(\theta)d\theta. \tag{71}$$

Under this new parametrization, the predictors' type which corresponds to the highest predictor's quality, τ_{dm}^{max}, is $\underline{\theta}^{dm}$ such that $\tau_{dm}^{max} = \cot^2(\underline{\theta}^{dm})$.

Moreover, determining the equilibrium stopping rule τ^* is equivalent to finding $\theta^* \in [\underline{\theta}^{dm}, \pi/2]$ which solves: $\exp(-\rho c) = G(\theta^*)$ where (see eq.(22) in the text):

$$G(\theta^*) \equiv \int_{\underline{\theta}^{dm}}^{\theta^*} r(\theta, \theta^*) \frac{h(\theta)}{1 - H(\underline{\theta}^{dm})} d\theta + \left(1 - \int_{\underline{\theta}}^{\theta^*} \frac{h(\theta)}{1 - H(\underline{\theta}^{dm})} d\theta\right), \quad \text{for } \theta^* \in \left[\underline{\theta}^{dm}, \frac{\pi}{2}\right], \tag{72}$$

with

$$r(\theta, \theta^*) = \frac{g(\theta, \theta^*)}{g(\theta^*, \theta^*)} = \left(\frac{\rho^2\sigma_\omega^2\sigma_\eta^2\tau(\theta^*) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}^2(\theta^*; \underline{\theta}^{dm})}{\rho^2\sigma_\omega^2\sigma_\eta^2\tau(\theta) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}^2(\theta^*; \underline{\theta}^{dm})}\right)^{\frac{1}{2}}, \tag{73}$$

where the expression for $r(\theta, \theta^*)$ follows from eq.(63) in the appendix of the paper.

B A family of distribution for predictors' types

Now consider the following family of distribution for predictors' types θ, indexed by $n \geq 1$ such that:

$$H(\theta) = \sin^{2n+1}(\theta), \quad \text{and therefore} \quad h(\theta) = (2n+1)\cos(\theta)\sin^{2n}(\theta). \tag{74}$$

For this family of distribution, the cumulative probability distribution of τ and its density are readily derived using eq.(69) and eq.(70) in Section III.A of this appendix. We obtain:

$$\Psi(\tau) = 1 - \frac{1}{(1+\tau)^{n+\frac{1}{2}}}. \tag{75}$$

and

$$\psi(\tau) = \frac{n+\frac{1}{2}}{(1+\tau)^{n+\frac{3}{2}}}. \tag{76}$$

Thus, $(1+\tau)$ has a power law distribution with parameters $(1, n+3/2)$. For this family of probability distributions, one can easily compute the average quality of data miners' predictors, $\bar{\tau}_{dm} \equiv \mathsf{E}_h\left[\tau(\theta')|\, \underline{\theta}^{dm} \leq \theta' \leq \theta^*\right]$. Indeed, as $\tau(\theta) = \cot(\theta)^2$, we have:

$$\begin{aligned}
\mathsf{E}_h\left[\tau(\theta')|\, \underline{\theta}^{dm} \leq \theta' \leq \theta^*\right] &= \frac{\int_{\underline{\theta}^{dm}}^{\theta^*} \frac{\cos^2(\theta)}{\sin^2(\theta)}(2n+1)\cos(\theta)\sin^{2n}(\theta)d\theta}{\sin^{2n+1}(\theta^*) - \sin^{2n+1}(\underline{\theta}^{dm})} \\
&= \frac{\int_{\underline{\theta}^{dm}}^{\theta^*}(2n+1)\cos^3(\theta)\sin^{2n-2}(\theta)d\theta}{\sin^{2n+1}(\theta^*) - \sin^{2n+1}(\underline{\theta}^{dm})} \\
&= \frac{\int_{\underline{\theta}^{dm}}^{\theta^*}(2n+1)(\cos(\theta)\sin^{2n-2}(\theta) - \cos(\theta)\sin^{2n}(\theta))d\theta}{\sin^{2n+1}(\theta^*) - \sin^{2n+1}(\underline{\theta}^{dm})} \\
&= \frac{\left[\frac{2n+1}{2n-1}\sin^{2n-1}(\theta) - \sin^{2n+1}(\theta)\right]_{\underline{\theta}^{dm}}^{\theta^*}}{\sin^{2n+1}(\theta^*) - \sin^{2n+1}(\underline{\theta}^{dm})}
\end{aligned} \tag{77}$$

In our numerical examples, we also assume that the cumulative distribution of the quality of experts' signals is also given by eq.(75) except for its lower bound that we set at $\underline{\theta}^{ex} = 0$. Thus, the average quality of experts' signals, $\bar{\tau}_{ex}$ is given by eq. (77) for $\theta^{dm} = 0$ (i.e., $\bar{\tau}_{ex} = \mathsf{E}_h\left[\tau(\theta')|\, 0 \leq \theta' \leq \theta^*\right]$. This specification is not key: One can use for the cumulative distribution of the quality of experts' signals any other distribution $H(.)$ in the family of distributions given in eq.(74).

To compute $G(.)$ (in eq.72) and solve for θ^* (and therefore data miners' stopping rule τ^*), we just need to compute $\bar{\tau}(\theta^*; \underline{\theta})$ and the integral in eq.(72). The expression for

$\bar{\tau}(\theta^*;\underline{\theta}^{dm})$ is given by

$$\begin{aligned}\bar{\tau}(\theta^*;\underline{\theta}^{dm}) &= \mu^*\bar{\tau}_{dm} + (1-\mu^*)\bar{\tau}_{ex} \\ &= (1-H(\theta^*))\,\mathsf{E}_h\left[\tau(\theta')|\,\underline{\theta}^{dm} \leq \theta' \leq \theta^*\right] + H(\theta^*)\,\mathsf{E}_h\left[\tau(\theta')|\,0 \leq \theta' \leq \theta^*\right],\end{aligned} \tag{78}$$

which is readily computed using eq.(77). Thus, we just need to explain how to compute:

$$\int_{\underline{\theta}}^{\theta^*} \frac{h(\theta)}{\left(\rho^2\sigma_\omega^2\sigma_\eta^2\cot^2(\theta) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}^2(\theta^*;\underline{\theta}^{dm})^2\right)^{\frac{1}{2}}} d\theta = (2n+1)D(0,n), \tag{79}$$

where

$$D(0,n) \equiv \int_{\underline{\theta}^{dm}}^{\theta^*} \cos(\theta)\sin^{2n+1}(\theta)\left(\rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}^2(\theta^*;\underline{\theta}^{dm})^2\sin^2(\theta)\right)^{-\frac{1}{2}} d\theta. \tag{80}$$

We now explain how to compute $D(0,n)$. To this end, let define $D(k,m)$ as

$$D(k,m) \equiv \int_{\underline{\theta}^{dm}}^{\theta^*} \cos(\theta)\sin^{2m+1}(\theta)\left(\rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*;\underline{\theta}^{dm})^2\sin^2(\theta)\right)^{k-\frac{1}{2}} d\theta.$$

Integrating by part, one obtains

$$\begin{aligned}D(k,m) = &\left[\frac{1}{2\left(k+1-\frac{1}{2}\right)}\sin^{2m}(\theta)\left(\rho^2\sigma_\omega^2\sigma_\eta^2\tau(\theta) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*;\underline{\theta}^{dm})^2\right)^{k+1-\frac{1}{2}}\right]_{\underline{\theta}^{dm}}^{\theta^*} \\ &- \frac{2m}{2\left(k+1-\frac{1}{2}\right)}D(k+1,m-1)\end{aligned} \tag{81}$$

and

$$D(k,0) = \left[\frac{1}{2\left(k+1-\frac{1}{2}\right)}\left(\rho^2\sigma_\omega^2\sigma_\eta^2\tau(\theta) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*;\underline{\theta}^{dm})^2\right)^{k+1-\frac{1}{2}}\right]_{\underline{\theta}^{dm}}^{\theta^*}$$

Hence, $D(0,n)$ can be expressed as a function of $D(1,n-1)$, which can be expressed as a function of $D(2,n-2)$ etc. until $D(n,0)$. Thus, one can obtain a closed-form expression for $D(0,n)$ and therefore $G(.)$ for any n.

C Special cases

The numerical examples used in the paper correspond to the cases $n = 1$ and $n = 2$ for the family of distributions presented in the previous section. In these two cases, we derive in Lemma 1 and 2 below closed form expressions for $\bar{\tau}(\theta^*, \underline{\theta}^{dm})$,and $G(\theta^*)$. We then use these expressions to numerically solve for θ^* (and therefore τ^*) and compute, for instance, the first moment of asset managers' expected trading profit ($\mathsf{E}[\bar{\Pi}(\tau(\theta))]$) in Section VI.A. The Mathematica code used for all numerical results in the paper is available upon request.

C.1 Case 1 .

Lemma 1. *When $h(\theta) = 3\cos(\theta)\sin^2(\theta)$, we have:*

$$\bar{\tau}(\theta^*, \underline{\theta}^{dm}) = (1-\sin^3(\theta^*))\frac{3[\sin(\theta^*) - \sin(\underline{\theta}^{dm})] - [\sin^3(\theta^*) - \sin^3(\underline{\theta}^{dm})]}{\sin^3(\theta^*) - \sin^3(\underline{\theta}^{dm})} + 3\sin(\theta^*) - \sin^3(\theta^*) \tag{82}$$

and

$$G(\theta^*) = 1 - \frac{\sin^3(\theta^*) - \sin^3(\underline{\theta}^{dm})}{1 - \sin^3(\underline{\theta}^{dm})} \tag{83}$$

$$+ \left(\rho^2\sigma_\omega^2\sigma_\eta^2\cot^2(\theta^*) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*, \underline{\theta}^{dm})^2\right)^{\frac{1}{2}} \frac{\delta(\theta^*) - \delta(\underline{\theta}^{dm})}{1 - \sin^3(\underline{\theta}^{dm})} \tag{84}$$

where

$$\delta(\theta) = \frac{\left(\rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*, \underline{\theta}^{dm})^2\sin^2(\theta)\right)^{\frac{1}{2}}}{\bar{\tau}(\theta^*, \underline{\theta}^{dm})^4}\left(\bar{\tau}(\theta^*, \underline{\theta}^{dm})^2\sin^2(\theta) - 2\rho^2\sigma_\omega^2\sigma_\eta^2\right). \tag{85}$$

Proof of Lemma 1. The expressions for $\bar{\tau}(\theta^*, \underline{\theta}^{dm})$ and $G(\theta^*)$ follows from the derivations in Section III.B for the general case.

C.2 Case 2.

Lemma 2. *When $h(\theta) = 5\cos(\theta)\sin^4(\theta)$,*

$$\bar{\tau}(\theta^*, \underline{\theta}^{dm}) = (1-\sin^5(\theta^*))\frac{\frac{5}{3}[\sin^3(\theta^*) - \sin^3(\underline{\theta}^{dm})] - [\sin^5(\theta^*) - \sin^5(\underline{\theta}^{dm})]}{\sin^5(\theta^*) - \sin^5(\underline{\theta}^{dm})} + \frac{5}{3}\sin^3(\theta^*) - \sin^5(\theta^*) \tag{86}$$

and

$$G(\theta^*) = 1 - \frac{\sin^5(\theta^*) - \sin^5(\underline{\theta}^{dm})}{1 - \sin^5(\underline{\theta}^{dm})} \tag{87}$$

$$+ \left(\rho^2\sigma_\omega^2\sigma_\eta^2\cot^2(\theta^*) + \rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*, \underline{\theta}^{dm})^2\right)^{\frac{1}{2}} \frac{\delta(\theta^*) - \delta(\underline{\theta}^{dm})}{1 - \sin^5(\underline{\theta}^{dm})} \tag{88}$$

with

$$\delta(\theta) = \frac{\left(\rho^2\sigma_\omega^2\sigma_\eta^2 + \bar{\tau}(\theta^*, \underline{\theta}^{dm})^2\sin^2(\theta)\right)^{\frac{1}{2}}}{\bar{\tau}(\theta^*, \underline{\theta}^{dm})^6}\left[\frac{8}{3}\rho^4\sigma_\omega^4\sigma_\eta^4 - \frac{4}{3}\rho^2\sigma_\omega^2\sigma_\eta^2\bar{\tau}(\theta^*, \underline{\theta}^{dm})^2\sin^2(\theta) + \bar{\tau}(\theta^*, \underline{\theta}^{dm})^4\sin^4(\theta)\right]. \tag{89}$$