Competing Models*

José Luis Montiel Olea[†] Pietro Ortoleva[‡] Mallesh Pai[§] Andrea Prat[¶]

First version: October 2017
This version: January 25, 2022

Abstract

Different agents need to make a prediction. They observe identical data, but have different *models*: they predict using different explanatory variables. We study which agent believes they have the best predictive ability—as measured by the smallest subjective posterior mean squared prediction error—and show how it depends on the sample size. With small samples, we present results suggesting it is an agent using a *low-dimensional* model. With large samples, it is generally an agent with a *high-dimensional* model, possibly including irrelevant variables, but never excluding relevant ones. We apply our results to characterize the winning model in an auction of productive assets, to argue that entrepreneurs and investors with simple models will be overrepresented in new sectors, and to understand the proliferation of "factors" that explain the cross-sectional variation of expected stock returns in the asset-pricing literature.

*We thank Sylvain Chassang, Kfir Eliaz, Ben Golub, Yuhta Ishii, Annie Liang, Jonathan Libgober, George Mailath, Stephen Morris, Wolfgang Pesendorfer, David Pearce, Luciano Pomatto, Rani Spiegler, Andrei Shleifer, Stefanie Stancheva, four anonymous referees, and the participants of many seminars and conferences for their useful comments and suggestions. We thank Stefano Giglio for helpful discussions and feedback. Dibya Mishra, Ken Teoh and Amilcar Velez provided excellent research assistance. Pai gratefully acknowledges the financial support of NSF Grant CCF-1763349.

[†]Department of Economics, Columbia University. Email: jm4474@columbia.edu
[‡]Department of Economics and SPIA, Princeton University. Email: pietro.ortoleva@princeton.edu.
[§]Department of Economics, Rice University. Email: mallesh.pai@rice.edu.
[¶]Columbia Business School and Department of Economics, Columbia University.
Email: andrea.prat@columbia.edu

1 Introduction

The value that individuals assign to a choice often depends on how well they believe they can predict unknown variables. How much an entrepreneur is willing to pay for a company, or whether they choose to enter a new market, depends on their belief in their own ability to predict and respond to future conditions, like market demand, costs, and competition. Households are more likely to invest in financial assets if they believe they can predict future market values.

In this paper, we study how individuals' assessments of their own predictive ability interact with the models they use and the available sample size. Our agents are Bayesian and observe the same data, but their predictions are based on different models: some agents believe only a few covariates matter for predictions, while others believe that many more do. We ask: What are the characteristics of the model of the agent who, after observing the data, believes they have the best predictive ability, as measured by the smallest subjective posterior mean squared prediction error (subjective MSPE)? Colloquially, a candidate who believes they have the best predictive ability may be described as the most *confident*. Similarly, if the subjective MSPE is below the (unknown) objective MSPE of their model, they may be described as *overconfident*. In what follows, we refer to the agent's assessment by subjective MSPE, but in our applications we expand on this confidence/ overconfidence interpretation to deliver novel implications to the behavioral literature on overconfidence.

We show that the answer depends on the model's dimension and the sample size. With *small samples*, agents with the smallest subjective MSPE use a *low-dimensional* model, using only a few covariates, regardless of the true data generating process (DGP). In contrast, with *large samples*, agents with the smallest subjective MSPE use a *high-dimensional model*, possibly including irrelevant covariates, but never excluding relevant ones. In single-agent decision problems, this results in novel comparative statics: the dimension of agents' models and the dataset's sample size influence the value they assign to each action, holding fixed other standard considerations (e.g., risk aversion, outside options, etc.). In settings where agents compete and relative subjective expected prediction error matters, model dimension and sample size determine the winning model.

Our model. As a concrete example, consider a second-price auction where a productive asset is sold to the highest bidder. The new owner of the asset will choose an action a, and her payoff will be given by $M - (a - y)^2$, where M is a known positive quantity and y is a random variable. Thus, the value of the asset depends on how well the agent can predict y.

There are multiple interesting economic issues in such a setting. Bidders may have

different payoff functions, different action sets, or different information. We abstract from all those issues and focus on the impact of using priors that involve a simpler (as compared to a more complex) relationship between explanatory variables and the variable of interest y. Specifically, suppose all agents agree that y is a linear function of a number of covariates $\{x_j\}_{j\in\{1,\dots,k\}}$ plus a noise term, i.e., $y = \sum \beta_j x_j + \epsilon$. Both the β_j's and the variance of ϵ are unknown, and agents may have different prior distributions on them. In particular, some agents may believe that only a subset of the covariates matters for predicting y.

All agents are given the same data: n independent draws of x and y, according to an unknown process. Agents are Bayesian but have different priors, as in Harrison and Kreps (1978) or Morris (1994). Thus, each agent computes a posterior distribution of β_j's and the variance of ϵ, and will use these posterior distributions to solve for their optimal action. Notice that there is no winner's curse in this setup, as winning the auction has no effect on the winner's posterior distribution. Therefore, the auction has the usual equilibrium in dominant strategies. In this equilibrium, each agent bids her expected value of the asset if she becomes the owner and gets to choose a. The winner is the agent with the lowest subjective MSPE. As everything else is equal, this is a competition among models. We ask: What are the characteristics of the model that, after observing the data, has the lowest subjective MSPE?

Note that there is a trivial reason why certain models may have lower subjective MSPE: their priors may contain less uncertainty about the world. The most extreme case is when an agent is dogmatic and she has a (right or wrong) deterministic model. That agent would believe she has no prediction error and would always bid M in the auction described above. To focus on more interesting effects, we prove all our results under the assumption that, absent data, all agents have the same expected loss.

Results. Our first result, Lemma 1, characterizes the subjective MSPE of an agent as a function of their prior and observed data. We prove a subjective/ Bayesian variant of a standard decomposition to show that subjective MSPE can be written as the sum of two components, which we term 1) *model fit*: the agent's posterior expectation of the variance of the regression residual, ϵ; and 2) *model estimation uncertainty*: the agent's degree of uncertainty about the coefficients in her regression model. Crucially, we show that the latter depends on the model's dimension. This implies that, while our Bayesian agents use their posteriors to compute the best action and do not explicitly care about the dimension of their model, the dimension affects their subjective MSPE.

This characterization has two immediate implications, depending on the size of the dataset. Our first set of results pertains to the case of small samples. Here we show that

model estimation uncertainty plays a critical role. While agents who use only a few covariates may have a lower model fit, they will also have a lower model estimation uncertainty, since they have fewer parameters to estimate. One complication with small samples is that the actual realized dataset matters, not just the agent's prior. Assuming that priors take a convenient conjugate form typical in Bayesian linear regression, we show how small sample sizes favor small models—even assuming that all agents have the same prior expectation about their prediction error. First, Proposition 1 shows that when the dataset consists of a single datapoint, the lowest subjective MSPE is achieved by a model that contains a single covariate (regardless of the realized data, the true DGP, or the parameters of the prior). Second, Proposition 2 shows that for any fixed sample size and for any true data generating process, lower-dimensional models have a lower subjective MSPE with high probability, as long as the prior on the variance of ϵ is high enough. Intuitively, with small samples, uncertainty about the parameters of the model plays a crucial role. Smaller models have an advantage since uncertainty about the parameters decreases faster as data accumulates. Even if these smaller models are misspecified, if the sample is small their model fit will not be much lower, meaning that they will have the lowest subjective MSPE. Third, we show that smaller models also have a smaller subjective MSPE when agents all believe they know the variance of the error and this variance is common. Finally, we show that this is also true when agents need to compute their future expected subjective MSPE before data is realized, but know that a data set of size n, for any n, is going to be revealed before the action is chosen.

Next, we consider the case of large sample size. Here, model estimation uncertainty vanishes: agents will have no uncertainty about their fitted parameters, even if they are using the wrong model. Subjective MSPE is therefore based solely on *model fit*. Proposition 3 then shows that models that omit a covariate that is relevant for prediction never prevail. At the same time, we also show that high-dimensional models—those that contain additional covariates irrelevant to the true DGP—may continue to win, even asymptotically. Even though these high-dimensional models will converge to the true DGP, for any finite sample they remain strictly different. We show that the probability of winning for high-dimensional models remains strictly above zero, even asymptotically. In turn, this shows that the role of priors does not vanish asymptotically: it continues to affect a model's probability of winning, even with arbitrarily large samples.

Applications. In the main body of the paper, we discuss the case of an auction of a productive asset as a leading example. In Section 6, we present two additional applications. First, we consider a simple model of entry in which returns depend on prediction error:

for example, the decision of an entrepreneur to enter a new sector, or the decision of a household to invest in a risky asset. Conditional on entry, the agent must make further choices. For example, as in classic organizational economics models, the entrepreneur must choose a strategy a that fits the (predicted) state of the world y and the loss function is the quadratic difference between a and y, as in, e.g., Marschak and Radner (1972), Milgrom and Roberts (1992) or Alonso et al. (2008). Alternatively, the investor must predict price movements to profitably buy/sell the asset. Agents may have simple or complex models: they make their forecasts using few or many covariates. The results of this paper provide a novel comparative static: in new sectors/asset classes, we should observe an overrepresentation of investors with simple models, even when reality is complex. We connect this to the literature on overconfidence of entrepreneurs and investors, and provide anecdotal evidence from cryptocurrencies.

Second, we use our framework to understand the proliferation of factors that explain the cross-sectional variation of expected stock returns in the asset-pricing literature. We argue that the increase in the number of test portfolios used to compute the popular Fama-French cross-sectional regressions mechanically favors asset-pricing models with several factors. Our empirical analysis on the evolution of the "factor zoo" can be viewed as a particular instance of a simple model of scientific progress. Early models (when there is little data available) may be overly simple relative to the truth, and become more complicated as samples accumulate.

Related Literature. Our results sit within the large and growing body of work in economic theory on agents with misspecified models; we defer a full discussion to Section 7. One key difference is that in most of the literature, misspecified models are evaluated using their objective performance. Our paper, along with a few contemporaneous or subsequent ones (Eliaz and Spiegler, 2018; Levy et al., 2019; He and Libgober, 2020), focuses instead on agents' *subjective* perception of their prediction error, the key metric in our applications.

Our results may also, at a high level, be reminiscent of model-selection methods in Statistics and Machine Learning, with one big difference: our results emerge as the outcome of competition among Bayesian decision makers using different models. By contrast, the model selection literature proposes and studies techniques to explicitly penalize high-dimensional models. The Bayesian statisticians in our paper cannot discard covariates.

Outline. The remainder of the paper is organized as follows. Section 2 outlines the formal model, and characterizes the subjective MSPE of a single agent, the foundation of our results. Section 3 illustrates the key trade-offs under competition with a simple numerical simulation. Section 4 then presents formal results for the case where the size of the dataset, n, is small,

while Section 5 considers the case where n is large. Section 6 studies the applications described above. Section 7 discusses the related literature, and Section 8 concludes.

2 Model and Single-Agent Problem

Agents want to predict a real-valued variable y. There are k real-valued covariates (or explanatory variables) $x \in \mathbb{R}^k$.

Data and Data Generating Process. Before making a prediction, agents observe a common data set, denoted D_n, composed of n i.i.d. draws of y and x. We denote the data as $D_n = (Y, X)$, where $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times k}$. The assumption that all agents observe the *same* data will be relevant for our applications (for example, in an auction setting, this avoids winner's curse).

A true Data Generating Process (DGP), denoted $\mathbb{P}$, determines the joint distribution of the random variables y and x. Most of our results assume only that true distribution of covariates has finite moments of all orders (and a positive definite matrix of second moments).

Statistical Models. Agents do not know $\mathbb{P}$ but work with a *statistical model*: a family of plausible joint distributions for y and x. In particular, agents posit a linear relation between y and the covariates $x \in \mathbb{R}^k$, i.e., conditional on x:

$$y = x'\beta + \epsilon, \qquad \text{where} \qquad \epsilon | x \sim \mathcal{N}_1(0, \sigma^2), \quad \beta \in \mathbb{R}^k, \tag{1}$$

that is, agents assume $y|x$ is a *homoskedastic linear regression with Gaussian errors* and parameters $(\beta, \sigma^2) \in \mathbb{R}^k \times \mathbb{R}_+$.[1,2]

Agents also assume that the covariates follow a distribution P which belongs to some parametric family.[3] This may or may not be the correct distribution. We assume that, for any element in this family, the matrix $E_P[xx']$ is positive definite and that the random vector

[1] Because covariates in x can be correlated, our framework allows the agents to consider a wide family of nonlinear relations. For example, the nonlinear process $y = 3\frac{x_1^3}{\sqrt{x_5}} + \epsilon$, can be accommodated by defining a new observable equal to $\frac{x_1^3}{\sqrt{x_5}}$. While not all nonlinear processes can be expressed this way, especially since we assume finitely many covariates, good approximations can always be achieved. Of course, once new variables are defined to obtain such approximations, there is no guarantee that all combinations lead to reasonable nonlinear models that other agents might hold.

[2] We use the notation $\mathcal{N}_k(\mu, \Sigma)$ to denote a multivariate normal distribution of dimension k with mean μ and covariance matrix Σ. See p. 171 of Hogg et al. (2006) for a textbook reference on this convention.

[3] A family of distributions $\mathcal{P}$ is said to be parametric if its elements are indexed by a finite-dimensional vector. One example is $x \sim \mathcal{N}_k(0, \Sigma)$, where Σ is an unknown positive definite matrix.

x has finite moments of all orders.[4]

Together, P, β, and σ^2 fully define a joint distribution over y and x, which we denote by Q_θ, with parameter $\theta := (\beta, \sigma^2, P)$.

Different Explanatory Variables. Different agents may consider different explanatory variables in x as relevant for their prediction. We assume that agents consider at least one explanatory variable in their models. The following notation will be useful. If $\{1, 2 \ldots, k\}$ label the explanatory variables in x, we denote by $J \subseteq \{1, 2 \ldots, k\}$ the subset that an agent considers possibly relevant for prediction. For a given vector β, the subvector consisting solely of the components in $J \subseteq \{1, \ldots, k\}$ is denoted by β_J. Let x_J be the analogous subvector of x, and X_J the corresponding submatrix of X.

Misspecification. An agent who considers the variables in the set J has the statistical model $\{Q_{\theta_J}\}_{\theta_J \in \Theta_J}$. Here Θ_J is the set of parameters corresponding to variables J, i.e., $\theta_J := (\beta_J, \sigma^2, P)$. This model is said to be *misspecified* if there is no $\theta_J \in \Theta_J$ for which $Q_{\theta_J} = \mathbb{P}$ (and it is correctly specified otherwise). In other words, a statistical model is misspecified if it does not contain the true DGP (Kleijn and Van der Vaart, 2012). When the true DGP $\mathbb{P}$ is also a Gaussian linear regression model as in (1) (which we will assume for some of our results), let J_0 denote the covariates with nonzero βs in the true DGP. Then, note that the model associated with any set of variables J for which $J_0 \not\subseteq J$ is necessarily misspecified.[5]

Priors. Agents are Bayesians. An agent who considers variables J as relevant for prediction has a prior π over Θ_J. It will be convenient to denote by $J(\pi)$ the set of variables that an agent with prior π considers relevant for prediction. Formally, let π_j denote the marginal distribution over β_j corresponding to prior π. If δ_0 denotes a Dirac measure at zero, then

$$J(\pi) := \{j \in 1, \ldots, k : \pi_j(\beta_j) \neq \delta_0\}.$$

In a slight abuse of terminology, we sometimes use $J \subseteq \{1, \ldots, k\}$ to refer to a *model*, which should be understood as the set of explanatory variables that are not exactly equal to zero

[4] Positive definiteness of the matrix rules out the case in which one covariate is a linear combination of some of the others.

[5] At first glance, it might seem reasonable to say that a model $y = \beta_1 x_1 + \epsilon$ need not be misspecified when the model is truly $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon'$, as long as the distributions of ϵ and $\beta_2 x_2 + \epsilon'$ coincide. But this is ruled out by the assumption that error distributions are restricted to be Gaussian, centered at zero, and independent of covariates.

under the prior π. Strictly speaking, though, a statistical model refers to the collection of distributions over data, given parameters as we have defined above; see McCullagh (2002).

Actions, Utility, and Optimal Prediction. Agents make a prediction of y given covariates x. Formally, they construct a prediction function f that maps x into y, i.e., $f : \mathbb{R}^k \to \mathbb{R}$. They minimize a standard quadratic loss function, equal to the square of the difference between the true y and their forecast f, i.e., $(y - f)^2$. Denote by $L(f, \theta)$ the agent's loss under prediction function f if the true DGP is Q_θ, i.e.,

$$L(f, \theta) := \mathbb{E}_{Q_\theta}[(y - f(x))^2]. \tag{2}$$

If π is the agent's prior over θ and D_n is the observed data, then characterizing the optimal prediction f^* is a standard problem. The agent chooses f to minimize $\mathbb{E}_\pi[L(f, \theta)|D_n]$, which can be rewritten as

$$\mathbb{E}_\pi[\sigma^2|D_n] + \mathbb{E}_\pi\left[\mathbb{E}_P[(x'\beta - f(x))^2] \mid D_n\right]. \tag{3}$$

The first term does not depend on f. The second term involves the average error incurred in predicting $x'\beta$ using $f(x)$.[6] With standard arguments (i.e., exchanging the order of integration and taking first-order conditions), we can see that this is minimized by

$$f^*_{(\pi, D_n)}(x) := x' \ \mathbb{E}_\pi[\beta|D_n] = x'_{J(\pi)}\mathbb{E}_\pi[\beta_{J(\pi)}|D_n]. \tag{4}$$

Thus, a Bayesian decision maker with a posterior $\pi|D_n$, model $J(\pi)$, and a square loss function, forecasts y at x as her Bayesian posterior mean of $x'\beta$. This is a standard result.

The agent's posterior loss, conditional on her using the optimal prediction function characterized above, is denoted $L^*(\pi, D_n)$. We refer to it as the *subjective posterior mean-squared prediction error* (subjective MSPE).

2.1 Decomposing Subjective MSPE

We now characterize an agent's subjective MSPE, our quantity of interest. The key forces at play will already be evident from the following lemma.

Lemma 1. *Suppose that $\beta_{J(\pi)}$ is independent of P under the posterior distribution. The*

[6] The inner expectation averages over values of x. The outer one averages over the values of β and P.

agent's subjective MSPE can be decomposed as:

$$L^*(\pi, D_n) = \underbrace{\mathbb{E}_\pi\left[\sigma^2 | D_n\right]}_{\textit{Model Fit}} + \underbrace{\operatorname{tr}\left(\mathbb{V}_\pi[\beta_{J(\pi)}|D_n]\ \mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}] \mid D_n\right]\right)}_{\textit{Model Estimation Uncertainty}}, \tag{5}$$

where $\mathbb{V}(\cdot)$ *is the variance-covariance operator, and* tr *is the trace operator.*

This lemma is reminiscent of standard decompositions of *mean-squared prediction error* in frequentist linear regression models (e.g., Hansen, 2021, Theorem 4.8), except that in this case it characterizes the *subjective* MSPE of the agent using their own prior. The lemma shows that the agent's subjective MSPE, $L^*(\pi, D_n)$, is the sum of two components. The first, the posterior expectation of σ^2_ϵ, is the agent's estimate of the irreducible noise in the system. We interpret this term as a measure of *model fit*, i.e., how well the model explains the data (as all unexplained variation must be ascribed to noise).

The second term, $\operatorname{tr}\left(\mathbb{V}_\pi[\beta_{J(\pi)}|D_n]\ \mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}] \mid D_n\right]\right)$, is the trace of the variance-covariance matrix of the coefficients of the model (adjusted by the posterior mean of $\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}]$). We interpret this term as a measure of how uncertain the agent is in her estimation of the parameters of the model according to her own prior, capturing *model estimation uncertainty*. To illustrate why, consider the simpler case in which the posterior mean of $\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}]$ is the identity matrix. Then, the second term reduces to $\operatorname{tr}\left(\mathbb{V}_\pi[\beta_{J(\pi)}|D_n]\right)$, i.e., $\sum_{j\in J(\pi)} \mathbb{V}_\pi[\beta_j|D_n]$; this is simply the sum of the posterior variances of the parameters β_j, indeed a measure of model estimation uncertainty. In the next section we will show that this decomposition has immediate implications for which model leads to the lowest subjective MSPE.

Remark 1. The independence of $\beta_{J(\pi)}$ and P under the posterior distribution will hold under general assumptions. For example, it holds when agents know the distribution of covariates, or if we consider statistical models in which P does not enter the parametric model of $y|x$ and (β, σ^2) does not affect the distribution of x.

3 Competing Models: A First Look

Suppose agents participate in a mechanism that selects the agent with the *lowest* subjective MSPE. A leading example, discussed in the introduction, is that of a second-price auction of a productive asset. In this scenario, the winner of the productive asset will choose an action a and her payoff will be given by $M - (a - y)^2$, where M is a known positive quantity and y is the variable agents aim to predict. Thus, the value of the asset depends on how well the agent can predict y. Assuming that different priors are the only dimension of heterogeneity among participants, and noting that there is no winner's curse, the auction will select the

agent with the lowest subjective MSPE. Applying Lemma 1, we immediately derive that this is the agent with *the best trade-off between model fit and model estimation uncertainty.*

Before we dive into formal results, we present a simple simulation to illustrate the key forces at play and our main findings. Suppose that there are six covariates, $\{x_1, \ldots, x_6\}$, of which only the first five are relevant for prediction in the true DGP, i.e., $y = \sum_{j=1}^{5} \beta_j x_j + \epsilon$, and $\epsilon | x \sim \mathcal{N}(0, \sigma^2)$. For simplicity, assume that each nonzero regression coefficient, β_j, is equal to 1, as is σ^2. Also assume that $x \sim N(0, \mathbb{I}_6)$ under the true DGP.

Considering all subsets of covariates, there are 63 agents with linear regression models, one for each nonempty subset of $\{x_1, \ldots, x_6\}$. By construction, 61 are misspecified, one has the exactly correct model, and one has a model of higher dimension compared to the true DGP. For the simulation, we assume that agents' priors belong to a well-behaved family, common in Bayesian linear regression, and parametrize them so that all agents have the same subjective MSPE before seeing any data. Specifically, the agents are assumed to have the following parametric model for their covariates: $x_J \sim \mathcal{N}_{|J|}(0, \Sigma_J)$, where Σ_J is an unknown, positive definite matrix. Further, we assume that the priors over the β,σ^2 and Σ_J belong to the Normal-Inverse-Gamma-Inverse-Wishart family of Definition 1 below. In this simulation we set $a_0 = 2$, $b_0 = 10$, and $\gamma_0 = .001$.

Figure 1 plots the frequency of the size of the model of the agent with the lowest subjective MSPE for datasets of size $n \in \{1, \ldots, 50\}$. Two patterns emerge.

First, when n is small, low-dimensional models tend to win despite being misspecified. In fact, when $n = 1$, we can see in Figure 1 that the winner is a model with a *single* covariate.

Second, as n grows large, misspecified models never win. At the same time, the high-dimensional model that includes the redundant variable x_6 continues to win with relative frequency that appears to converge to a steady state close to 0.3—strictly above 0.

We will now show that both patterns hold more generally.

4 The Winner with Small n

We begin with the case in which the number of observations n is small. For tractability, we focus on a special class of priors, widely used in Bayesian linear regression.

Definition 1. A prior has the *Normal-Inverse-Gamma-Inverse-Wishart form* with *hyper-*

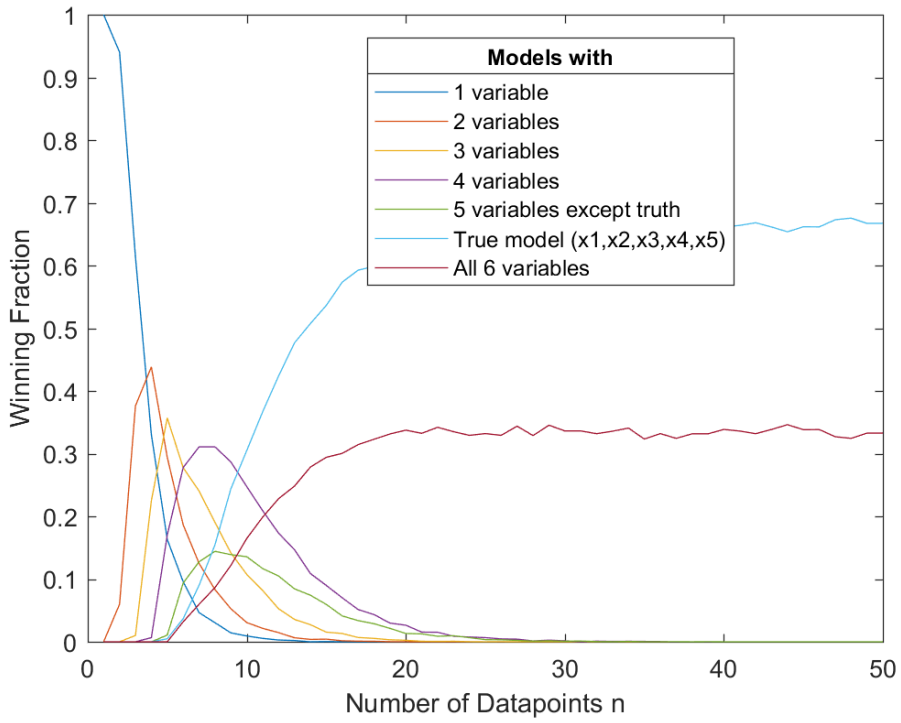

Figure 1: Winning probabilities on $5,000$ simulated datasets of size $n = 1, \ldots, 50$.

parameters (a_0, b_0, γ_0), with $(a_0, b_0, \gamma_0) \gg 0$ and $a_0 > 1$, if

$$\beta_J | \sigma^2 \sim \mathcal{N}_{|J|}\left(\mathbf{0}, \frac{\sigma^2}{\gamma_0 |J|} \mathbb{I}_{|J|}\right), \qquad \sigma^2 \sim \text{Inv-Gamma}(a_0, b_0),$$
$$x_J \sim \mathcal{N}_{|J|}(0, \Sigma_J), \qquad \Sigma_J \sim \text{Inv-Wishart}(\gamma_0 |J| \mathbb{I}_{|J|}, 2|J| + 1).$$

where Inv-Gamma(a_0, b_0) is the Inverse-Gamma distribution with parameters a_0 and b_0, and Inv-Wishart$(\gamma_0 |J| \mathbb{I}_{|J|}, 2|J| + 1)$ is the Inverse-Wishart distribution with $J \times J$ scale matrix $\gamma_0 |J| \mathbb{I}_{|J|}$ and degrees of freedom $2|J| + 1$.[7] The prior on Σ_J is assumed independent of (β, σ^2).

The Normal-Inverse-Gamma-Inverse-Wishart priors are conjugate priors for the Gaussian linear regression model and the posterior can be expressed in a closed form as a function of the data—see Appendix A.3 for details. All results in this section have analogs in the setting where the distribution P on covariates is assumed to be known (not necessarily Gaussian) and $E_P[xx'] = \mathbb{I}_k$.

[7] The Inverse-Gamma is a two-parameter family of distributions on the positive real line. For parameters $a_0 > 1, b_0 \geq 0$ it has mean $b_0/a_0 - 1$. The Inverse-Wishart is a two-parameter family of distributions on real-valued positive definite matrices. The first parameter (scale matrix) is a symmetric positive definite matrix. The second is a nonnegative scalar at least as large as the dimension of the scale matrix.

Using this family of priors allows us to further simplify the expression for subjective MSPE in Lemma 1.

Lemma 2. *Consider a prior π such that the implied distribution over $(\beta_J, \sigma^2, \Sigma_J)$ are as in Definition 1. Then*

$$L^*(\pi, D_n) = \underbrace{\mathbb{E}_\pi[\sigma^2|D_n]}_{\textit{Model Fit}} + \underbrace{\mathbb{E}_\pi[\sigma^2|D_n]\left(\frac{|J(\pi)|}{n+|J(\pi)|}\right)}_{\textit{Model Estimation Uncertainty}}. \tag{6}$$

One convenient feature of the family of priors in Definition 1 is that it implies that all agents with priors in this family with the same hyperparameters have the same subjective MSPE when no data is released. This is readily checked by substituting $n = 0$ in (6). Differences in subjective MSPE arise therefore only from the fact that the subjective MSPE evolves differently for models of different dimensions.

Equation (6) shows that the dimension of an agent's model enters explicitly into the subjective MSPE via the term $|J(\pi)| \,/\, (n + |J(\pi)|)$. As this is increasing in $|J(\pi)|$, higher-dimensional models have a disadvantage: since they have more parameters to estimate, their model estimation uncertainty will decrease more slowly. For higher-dimensional models to have lower subjective MSPE, therefore, they must compensate for this by a sufficiently better model fit. Specifically, the ratio between the model fits must be such that

$$L^*(\pi, D_n) \geq L^*(\pi', D_n) \iff \frac{\mathbb{E}_\pi[\sigma^2|D_n]}{\mathbb{E}_{\pi'}[\sigma^2|D_n]} \geq \frac{\left(1 + \frac{|J(\pi')|}{n+|J(\pi')|}\right)}{\left(1 + \frac{|J(\pi)|}{n+|J(\pi)|}\right)}. \tag{7}$$

In general, however, characterizing the model fit is analytically difficult, as it depends also on the realized data D_n, which in small samples can vary substantially. The following results describe how lower-dimensional models have lower subjective MSPE in some cases in which model fit is simple to analyze. First, it holds true when the sample size $n = 1$; we show this to be the case for a model using a single covariate. Second, it holds true when prior mean of σ^2 is high enough relative to n. Third, it applies when either all agents believe they know the variance of the error term, σ^2, and have the same belief, or when subjective MSPE is computed before actual data is released but the agents know that it will be released before they must choose their action—a case that has direct applications (as we will discuss later).

4.1 One-Dimensional Models Win when $n = 1$

The first result shows that when $n = 1$, low-dimensional models have the lowest subjective MSPE, regardless of the realized data, the true DGP, or the parameters of the prior.

Proposition 1. *Suppose all priors are as in Definition 1 with shared hyperparameters. Suppose also that for every single covariate model, i.e., every J such that $|J| = 1$, there is an agent with that model, and that all agents use at least one covariate. If $n = 1$, then the agent with lowest subject MSPE is an agent with a single covariate model.*

We prove this result by showing that, when $n = 1$, model fit is minimized by some model that considers only a single variable. This means that there is no model with more than one covariate that can improve the model fit of the best one-dimensional model.

4.2 Low-Dimensional Models Win when Prior Mean on σ^2 Is High

The sharp characterization obtained for $n = 1$ does not hold for other sample sizes. Indeed, our simulations show that models with more than one covariate may have the lowest subjective MSPE with positive probability for every $n > 1$. As we have discussed, the identity of this model depends on the trade-off between model fit and model estimation uncertainty.

One way to capture the advantage of small-dimensional models when n is small is the following. One characteristic of few data points is that the prior continues to play a relevant role. This can be captured by making sure that the prior mean over σ^2 (which equals $b_0/(a_0 - 1)$) is large enough relative to the amount of data. In this case, high-dimensional models cannot improve sufficiently on the model fit term relative to lower-dimensional models (as the model fit is poor for any model). Therefore, in this setting also, the advantage that low-dimensional models have in terms of model uncertainty is the dominant factor.

Proposition 2. *Let Π be a finite set of agents' priors that satisfy Definition 1 with shared hyperparameters (a_0, b_0, γ_0). Let $|\underline{J}|$ be the size of the smallest model in this set. Fix the size of the dataset n. For any $p \in (0, 1)$, there exists b_0 large enough so*

$$\mathbb{P}\left(D_n : \exists \pi^* \in \underset{\pi \in \Pi}{\operatorname{argmin}}\, L^*(\pi, D_n) \text{ s.t. } |J(\pi^*)| = |\underline{J}|\right) > p,$$

i.e., with probability at least p over datasets D_n, the agent with the lowest subjective MSPE has the smallest size model among all the agents.

As an illustration of this result, let us return to the simulations in Section 3. Figure 2 reports the winning fraction of models of size 1, as we increase the shared hyperparameter

b_0 (all other simulation parameters stay the same as above). Growing b_0 corresponds to a larger prior mean for all agents.

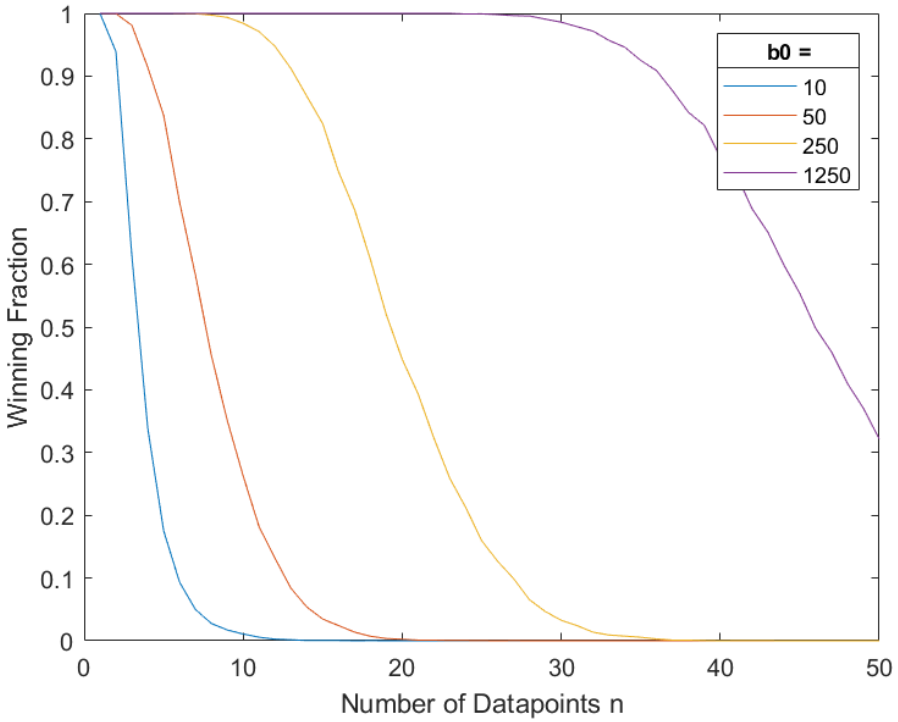

Figure 2: Winning rates for models with one covariate as the shared hyperparameter b_0 increases. All other simulation parameters are the same as in Figure 1.

4.3 Low-Dimensional Models Win when σ^2 Is Known or Data Are Not Released but Expected

We conclude this analysis by considering two other cases in which the model fit is easy to solve analytically, which again show an advantage of lower-dimensional models. We present them as two observations since they follow directly from Lemma 2 above.

Observation 1. Suppose all agents treat σ^2 as a known common value, but the priors on $\beta|\sigma^2$ and Σ are as in Definition 1. Then, for any D_n,

$$L^*(\pi', D_n) < L^*(\pi, D_n) \iff |J(\pi')| < |J(\pi)|.$$

That is, if σ^2 is believed to be known (possibly incorrectly), subjective MSPE is ranked by model dimension.

This result shows that when all agents believe they know the error variance σ^2, then model dimension induces a precise ranking between models: smaller models *always* have smaller subjective MSPE. This result follows directly from Lemma 2.

We now turn to the case in which agents have not received any data, but know that they will receive n data points before making their choice of action. They therefore have to compute their expected subjective MSPE, which we denote by $\mathbb{E}_{\pi'}[L^*(\pi', D_n)]$.

Observation 2. For any π and π',

$$\mathbb{E}_{\pi'}[L^*(\pi', D_n)] < \mathbb{E}_{\pi}[L^*(\pi, D_n)] \iff |J(\pi')| < |J(\pi)|.$$

This result shows that if agents have not received any data but know that they will receive it later, again we find small-dimensional models have lower expected MSPE. This observation also follows straightforwardly from Lemma 2: by the martingale property of beliefs, the expected model fit is constant among models, meaning that the winner must be a low-dimensional model.

5 The Winner with Large n

We now characterize the winner for large n. Our results will be derived for a much more general class of priors than the previous section, but it is helpful to start by recalling Lemma 2, which assumes Normal-Inverse-Gamma-Inverse-Wishart priors. In this case, the subjective MSPE is

$$\underbrace{\mathbb{E}_{\pi}[\sigma^2|D_n]}_{\text{Model Fit}} + \underbrace{\mathbb{E}_{\pi}[\sigma^2|D_n]\frac{|J(\pi)|}{n+|J(\pi)|)}}_{\text{Model Estimation Uncertainty}}.$$

From this formula, it is immediate to see that model estimation uncertainty vanishes as n grows large, making the model fit the crucial aspect. This means that, because misspecified models have worse model fit than correctly specified ones, they must therefore also have worse subjective MSPE when n is large enough.

The comparison is, however, less straightforward between a model that uses the exact same variables as the true DGP and another that also includes additional irrelevant covariates. For both, model fit converges to the true residual variance (σ_0^2), since both are correctly specified, and model estimation uncertainty converges to zero. Which one has lower subjective MSPE depends on how quickly these converge, which in turn depends on the realized data and on the prior. Our earlier simulations suggest that the long-run behaviors may be such that the larger model may continue to win, with a probability bounded away from zero

even at the limit. We will now show how this holds in general.

For this analysis, we do not need to assume that priors have a specific form as we did in the previous section. We simplify our analysis in the body of the paper by making two assumptions: that the true DGP is of the linear Gaussian form, which allows some models to be identical to the true DGP (Assumption 1); and that priors over the β_is have full support with a smooth density (on the subset of relevant covariates $J(\pi)$), while priors on σ^2_ϵ are not degenerate (Assumption 2).

Assumption 1. *There exists a parameter* $\theta_0 := (\beta_0, \sigma_0^2, P_0)$ *such that* $Q_{\theta_0} = \mathbb{P}$.

Assumption 2. *Priors are characterized by a smooth and strictly positive probability density function* $\pi(\cdot)$ *over* $({\beta_{J(\pi)}}', \sigma^2)' \in \mathbb{R}^{|J(\pi)|} \times \mathbb{R}_+$. *The prior over* $({\beta_{J(\pi)}}', \sigma^2)'$ *is independent of the prior over* P.[8] *In addition, for each agent, there exists* n *large enough for which* $\mathbb{E}_\pi[\sigma^2|D_n] < \infty$ *almost surely.*

Recall from the discussion on model misspecification in Section 2 that J_0 denotes the set of covariates that are relevant in the true DGP.

Proposition 3. *Suppose the true DGP* $\mathbb{P}$ *satisfies Assumption 1 with parameter* $(\beta_0, \sigma_0^2, P_0)$ *and associated subset of covariates that are relevant for prediction,* J_0. *Let* Π *be a finite collection of priors that satisfy Assumption 2. Further suppose that* Π *contains at least one prior* π^* *with* $J_0 \subseteq J(\pi^*)$. *If*

$$tr\left(n\mathbb{V}_\pi(\beta_{J(\pi)}|D_n)\mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}] \mid D_n\right]\right) = O_{\mathbb{P}}(1), \tag{8}$$

for every prior $\pi \in \Pi$, *then*

$$\lim_{n\to\infty} \mathbb{P}\left(\exists \pi \in \operatorname*{argmin}_{\pi\in\Pi} L^*(\pi, D_n) \ s.t \ J_0 \nsubseteq J(\pi)\right) = 0.$$

Moreover, for any π *for which* $J_0 \subset J(\pi)$,

$$\lim_{n\to\infty} \mathbb{P}\big(L^*(\pi, D_n) < L^*(\pi_0, D_n)\big) \in (0, 1],$$

where π_0 *is any prior for which* $J(\pi_0) = J_0$.

A crucial assumption in Proposition 3 is (8). This assumption will be verified whenever the posterior variance of β decreases to zero at rate n. Lemma 2 already tells us that this condition is satisfied in the special case of Normal-Inverse-Gamma-Inverse-Wishart priors.

[8] By definition, the prior of an agent for any β_κ, $\kappa \notin J(\pi)$, is degenerate at 0.

In fact, this condition holds very generally due to the Bernstein-von Mises theorem, which states that posterior distributions based on parametric models (misspecified or correctly specified) will typically behave like Gaussian distributions, with a variance that decreases at rate n.[9]

Proposition 3 has two takeaways. The first part tells us that, because it excludes relevant variables, a misspecified model never wins as the sample size grows large. Any model that is not misspecified will have lower subjective MSPE with probability approaching 1 as n grows large. The second part shows that *any* model larger than the true one defeats the latter with a probability that is *strictly positive*, even asymptotically.

We have already discussed the intuition for the first result. The assumptions of the theorem, i.e., Assumptions 1 and 2 combined with (8), guarantee that model estimation uncertainty converges to zero for all agents.

For the second result, from a technical perspective, our result is based on an asymptotic expansion for the posterior mean of the variance parameter in the linear regression model derived from the general results in Kass et al. (1990). This is a fairly technical result, so, for some intuition, let us return to the case of Normal-Inverse-Gamma-Inverse-Wishart priors. Here, when n is large, it is possible to approximate its distribution using asymptotic theory.

To this end, let $\widehat{\beta}_J$ denote the OLS estimator based on the variables listed in J, and let $\widehat{\sigma}^2_J$ be the corresponding residual variance estimator:

$$\widehat{\sigma}^2_J \equiv (Y - X'_J\widehat{\beta}_J)'(Y - X'_J\widehat{\beta}_J)/n.$$

A key observation in our analysis—which holds for Normal-Inverse-Gamma-Inverse-Wishart priors, but also for more general priors—is that as the sample size grows large,

$$n\left(E_{\pi_0}[\sigma^2|D_n] - E_{\pi}[\sigma^2|D_n]\right) = n\left(\widehat{\sigma}^2_{J_0} - \widehat{\sigma}^2_J\right) + O_{\mathbb{P}}(1), \tag{9}$$

where $O_{\mathbb{P}}(1)$ refers to a term that is bounded with high probability under $\mathbb{P}$. In the case of Normal-Inverse-Gamma-Inverse-Wishart priors, algebraic manipulations can be used to verify the approximation with a leading term equal to

$$-\gamma_0\beta'_0\beta_0(|J| - |J_0|).$$

Deriving an analogous result for other priors requires additional effort, given the lack of

[9]See the Bernstein-von Mises theorem for misspecified parametric models of Kleijn and Van der Vaart (2012). This result can be thought of as richer versions of the classical results concerning posterior distributions of misspecified models in Berk (1970).

closed-form solutions for the posterior distributions.[10]

Assumption 1, along with standard results from regression analysis—e.g., Equation 5.28 in Greene (2018) and Theorem 5.1 therein—implies $n(\widehat{\sigma}^2_{J_0} - \widehat{\sigma}^2_J)/\sigma^2_0$ converges in distribution to a chi-squared random variable with $|J| - |J_0|$ degrees of freedom. This means that the probability that the larger model wins can be approximated by the probability of the event:

$$\chi^2_{|J|-|J_0|} - \gamma_0(\beta_0'\beta_0/\sigma^2_0)(|J| - |J_0|) > (|J| - |J_0|).$$

In the context of our simulations—where the true DGP only included the first five covariates, with coefficients $\beta = (1, 1, 1, 1, 1)'$, and $\sigma^2 = 1$—the probability that a model with six variables defeats the true DGP is roughly

$$P(\chi^2_1 > 1 + 5\gamma_0).$$

When $\gamma_0 = .001$, this probability is 0.3161, which is close to what we see in Figure 1. We provide a more general formula in Appendix B.1.

It is important to remark that these results hold even if the true DGP is different from (1). For example, the distribution of errors in the true DGP may be heteroskedastic or non-normal with thin-enough tails, the distribution on covariates P may be misspecified as long as it has finite second moments, and the true DGP may not be linear. Although Proposition 3 is presented under special conditions (Assumptions 1 and 2), it is possible to prove that our results continue to hold under much weaker ones.[11] We hope that the simpler framework helps the reader understand the main forces at play in the competition among models.

Connection with the Akaike Information Criterion. A different way to understand our results is to relate the model selection they induce to the Akaike Information Criterion (AIC), a well-studied model selection criterion in econometrics and statistics. In what follows, we illustrate that the loss function of an agent with Normal-Inverse-Gamma-Inverse-Wishart prior is "close" to the AIC for the linear regression model.

Definition 2 (Akaike Information Criterion)**.** Given a dataset $D_n = (Y, X)$ with n data

[10] We refer the reader to Lemma 3, which uses the Kaas-Tierney-Kadane expansions of posterior moments in Kass et al. (1990) to verify the approximation.

[11] For example, for the expansion of Lemma 3 to hold, priors do not need to be smooth. It is sufficient that they are differentiable up to the fourth-order.

points and k possible covariates, the AIC for linear regression evaluates a model J as

$$L_{\text{Akaike}}(J, n, D_n) = \ln \widehat{\sigma}_J^2 + \frac{2|J|}{n},$$
$$\text{where } \widehat{\sigma}_J^2 = \frac{1}{n} \min_{\beta \in \mathbb{R}^{|J|}} (y - X_J\beta)'(y - X_J\beta).$$

The expression $\widehat{\sigma}_J^2$ is the OLS estimator of the residual variance based on a model with covariates X_J in the dataset D_n. As is well understood, the model with lower estimated variance may not be the model with the best out-of-sample performance. This is because selecting based on average residuals favors models that have more covariates, which may overfit the data. The AIC compensates for this by adding a penalty term equal to $\frac{2|J|}{n}$, i.e., twice the ratio of the number of covariates in the model and the number of data points. Algebra shows that, if agents have an uninformative Normal-Inverse-Gamma-Inverse-Wishart prior, then the posterior loss is approximately equal to

$$\ln\left(\widehat{\sigma}_J^2\right) + \ln\left(1 + \frac{|J|}{|J| + n}\right).$$

Thus, if the sample size is large and the agents' distribution of covariates is well-specified, the posterior loss of an agent with prior π is approximately equal to the AIC (with a penalty of $\ln(1 + |J|/(|J| + n)) \approx |J|/n$ instead of $2|J|/n$).

The prevalence of larger models in the model competition can be then associated to the "conservativeness" of the AIC for model selection. Proposition 3, however, makes clear that the relation is only qualitative: larger models will prevail in large samples, but the probability of a larger model being selected will continue to be affected by the prior.

Finally, it is worth reiterating that the foundations of the AIC are normative: the criterion was proposed as a way to select models to avoid overfitting. Conversely, our analyis provides a *positive* foundation for a solution similar to the AIC: we study the outcomes when Bayesian agents compete in a way that selects the agent with the lowest subjective MSPE.

6 Applications

In previous sections, we considered the auction of a productive asset as a leading example. We now discuss two additional applications. The first is a simple model of entry when returns depend on prediction error (e.g., investing). The results of this paper provide a novel comparative static: in new sectors/asset classes, we should observe an overrepresentation of investors with simple models, even when reality is complex. Second, we use our framework

to understand the proliferation of "factors" in the asset pricing literature.

6.1 Selection of Simple Models in Entry and Investment

We begin with an application to a single-agent decision problem. An agent is faced with a risky entry choice—she has to choose between a risky option and a safe one that gives her a utility normalized to 0. The utility of the risky option depends (in part) on the agent's ability to predict an unknown variable and take an action. This is the case of an entrepreneur who has the option to invest in a new venture, where expected returns depend in part on the ability to predict and adapt to future market demand, political situations, or trade agreements. Alternatively, this could be an individual investor considering trading an asset: returns depend on the ability of the investor to predict future price movements and trade accordingly.

Formally, suppose that the expected utility of the risky option is

$$\mathbb{E}(r) = v - L^*(\pi, D_n),$$

where v summarizes agent-specific costs and benefits that are independent of prediction error, while $L^*(\pi, D_n)$ is the component that depends on prediction error—in line with our notation, the subjective MSPE. The prior π summarizes the agent's prior belief about the relationship between the unknown variable they need to predict (e.g., market demand for an entrepreneur, price movement for an investor) and various observables they consider relevant. The agent's model is the set of variables they consider relevant for prediction. The data D_n is past data about this relationship. The agent knows v and π, observes D_n, and then chooses the risky option if its expected utility is positive.

The results of this paper directly apply. *Ceteris paribus* (fixing v), with few data points, agents with "simple models" are systematically more confident in their prediction error ($L^*(\pi, D_n)$ is lower) and are, therefore, more likely to take the risky option. Crucially, this holds whether their simple models are correct or not. This has an immediate implication that provides a novel comparative static: entrepreneurs with simple models are over-represented in sectors where little data has accumulated, even when the true DGP is complex. For example, this margin of selection suggests that, *ceteris paribus*, the entrepreneurs more eager to invest in a country that just opened up to foreign investment, or in new technologies, will tend to be those that believe they can predict future conditions using relatively few covariates—that have a simple model—even when reality is much more complex. Similarly, investors with simpler models are more likely to enter into new asset classes.

So far we have assumed that no new data is revealed after the investment decision. In

reality, however, new data accumulates after the initial investment/ choice to enter is made. Entrepreneurs or investors may take this into account in their decision, expecting to be able to refine their prediction. The discussion of Section 4.3 suggests that this only strengthens the selection in favor of simple models. To illustrate, consider the setup above but suppose that entrepreneurs must decide whether or not to invest before any data is revealed, but knowing that some data will be revealed at a later stage. As we discussed in Section 4.3, agents with simple models are more confident about how much they will be able to learn from the yet-to-be-released data. In this case, entrepreneurs/ investors with overly simplistic models will *always be over represented*.

Connections to Overconfidence. These findings connect to established empirical facts on overconfidence and entry. Several studies have shown that entrepreneurs are, by various measures, overconfident (Koellinger et al., 2007; Cooper et al., 1988). Similarly, a large body of evidence shows that (especially retail) investors are often overconfident about their knowledge and information (Odean, 1999; Statman et al., 2006). This is commonly attributed to either incorrect beliefs, with selection favoring individuals with overly optimistic priors, or post-decision bolstering.

Our results suggest a novel margin of selection related to model complexity in relation to overconfidence: entry into areas with limited past data (e.g., "new" areas) are systematically biased towards entrepreneurs or investors with models that are "simple." As we have seen, when the true DGP is complex, these individuals are also *overconfident* in their predictive ability: their subjective MSPE is on average lower than it should be. The relevant margin of selection may be the simplicity of the model, which generates both a higher likelihood of entrance and overconfidence in predictive ability. Similarly, investment in new asset classes is more common for investors who, *ceteris paribus*, have simple predictive models of price movement. Our results also suggest that this margin of selection is attenuated as data accumulates: entry into more established sectors, technologies, or countries may be systematically different from entry into "new" areas in terms of complexity of the entrepreneur's model; investors in established assets/ areas may be systematically different from investors in new asset classes/ trends (e.g., cryptocurrency).

Existing studies on overconfidence note also that agents often misreact, or underreact, to new information. For example, Odean (1999) studies the trading patterns of retail traders, and argues that their trades are systematically incorrect: "while investors' overconfidence in the precision of their information may contribute to this finding, it is not sufficient to explain it. These investors must be systematically misinterpreting information available to them. They do not simply misconstrue the precision of their information, but its very meaning."

This is consistent with our finding that, when the true DGP is complex and involves many variables, entry into trading may favor agents with incorrect models, i.e., ones that exclude covariates that are relevant for prediction and/or include irrelevant ones.

Anecdotal Evidence. To illustrate our novel comparative statics, we provide some anecdotal evidence related to cryptocurrency investment. Empirical evidence suggests that investors overconfident about their "investment literacy" are more likely to hold cryptocurrency (Kim and Hanna, 2021).[12] While such overconfidence can have several explanations, we think that the use of simple models in an environment with limited data may be a contributing factor.

Consider, for example, the "Stock-to-Flow model" of bitcoin pricing, published on the online platform *Medium* in March 2019.[13] This model posits that the current market price of bitcoin can be determined as a ratio of the current stock of bitcoin (i.e., the total amount that has been mined to date) and the flow (the amount of "new" bitcoins that miners are currently mining). The price predictions are implemented by regressing the logarithm of bitcoin price on the logarithm of the stock-to-flow ratio. The model was first estimated using monthly data: 111 data points from December 2009 to February 2019. The model is surely simplistic, but resulted in substantial buzz on social media.[14]

Since then, as data has accumulated (including higher frequency data), several more complicated models have been proposed. These range from traditional time series forecasting techniques (Munim et al., 2019) to support vector machines, neural networks, and random forests (Chen et al., 2020). This is consistent with our prediction that, as data accumulates, more complicated models are adopted.[15]

[12]The 2018 National Financial Capability Study Investor survey contains several questions that can be associated to "objective investment literacy" and "subjective investment literacy" (self-reported investment literacy). Their difference relates to overconfidence.

[13]`https://medium.com/@100trillionUSD/modeling-bitcoins-value-with-scarcity-91fa0fc03e25`

[14]As evidence of this, the number of Twitter followers of the author of this model (`@TrillionUSD`) grew from 5,000 in Feb, 2019 to 15,000 in June 2019, and it is currently over 1.7 million.

[15]The use of overly complicated models as data accumulates has been observed more broadly in finance. For example, consider the following quote from Campbell Harvey in Barberis et al. (2015):

> Overfitting is when you propose an overly complicated model to explain something rather simple ... This happens because the model does not adequately describe the effect but rather the noise in the data ... This type of data mining has been on the increase with today's easy availability of both financial data and computing power ...

While this point was made in the context of data mining and false-discovery in financial models, we note that such observed phenomena are consistent with our model.

6.2 Competing Factor Models[16]

A large body of work in finance has studied the cross-sectional variation in asset expected returns (i.e., why different assets earn different average returns). The classical asset-pricing framework (Jensen et al., 1972; Fama and MacBeth, 1973) posits that, at each point in time, asset returns are governed by a *multi-factor* model. The return of each asset—an individual stock or a portfolio—is an asset-specific linear combination of these factors (with time-invariant coefficients) plus random noise.

The search for factors to explain the cross-sectional variation of expected stock returns has produced hundreds of potential candidates. The literature has evolved from the parsimonious model of Fama and French (1993) using only three factors (market return, size premium, value premium) to the factor library in Feng et al. (2020) that contains 150 risk factors.[17]

We use our framework to understand the proliferation of factors in this literature. We argue that the increase in the number of "test portfolios" (i.e., portfolios of stocks that are formed by sorting on different characteristics; see Appendix C.2 for further details) used to compute the Fama-French cross-sectional regressions mechanically favors asset-pricing models with several factors.

To make this point, we view different collections of factors as competing models or, more precisely, competing sets of risk factors—a terminology that has been used, incidentally, by Fama and French (1993, 2015).[18] We then take the number of test portfolios as the number of available data points to predict the cross-section of expected returns. We consider the different factor models as different Bayesian agents competing to predict cross-sectional returns. Our results suggest that the winning model depends crucially on the sample size. With a few test portfolios, lower-dimensional factor models will be selected. Conversely, increasing the number of test portfolios favors high-dimensional factor models. In particular, Proposition 3 suggests that the subset of the factor library that minimizes subjective MSPE will include too many factors relative to the true best linear predictor of expected returns based on all the factor exposures to the 150 available risk factors.

Let i index an asset in the cross-section. The outcome variable Y_i denotes the excess return of asset i averaged over the different time periods for which data on returns and

[16] We thank Stefano Giglio for useful discussions in writing this section.

[17] See Appendix C.2 for a description of these factors, how they are constructed, and the year in which they were published.

[18] From Fama and French (1993, p. 12): "The average excess returns on the portfolios that serve as dependent variables give perspective on the range of average returns that *competing sets of risk factors* must explain." On p. 13: "The wide range of average returns on the 25 stock portfolios, and the size and book-to-market effects in average returns, present interesting challenges for *competing sets of risk factors*." Finally, in Fama and French (2015) "estimate the proportion of the cross-section of expected returns left unexplained by *competing models*" (emphasis added).

factors are available.[19] Each asset i has an associated 150-dimensional vector of covariates, X_i, containing the asset's factor loadings.[20] In principle, assets could be individual stocks or portfolios. We follow the literature and focus on portfolios.[21] The number of portfolios under consideration (N) gives the sample size for the cross-sectional regressions. We start with the 25 (5×5) portfolios sorted by size and book-to-market ratio as used in Fama and French (1993).[22] We then consider the larger collection of 2,875 (5×5) bivariate-sorted portfolios in Feng et al. (2020).[23]

Competing (Factor) Models. We consider "competition" between three different models. The first (Fama-French) uses only the original Fama-French factors (excess market return, and the "small minus big" and "high minus low" factors). The second model (Factor Zoo) uses all 150 factors in Feng et al. (2020). The third (FGX-DS) uses the 135 factors introduced up until 2011, plus five additional factors obtained by the "Double Selection" approach of Feng et al. (2020).[24] For each model, we consider the same set of hyperparameters (a_0, b_0), chosen to maximize the marginal likelihood of the largest model with the largest dataset, and set γ so that all models have the same prior subjective MSPE before any data. Details are provided in Appendix C.1.

Figure 3 presents our results. It shows the subjective MSPE of each model. To make units easier to interpret, the results are presented as a percentage relative to the worst competing model (out of the three under consideration).

Consistent with our theorems, if we consider only the 25 (5×5) bivariate sorted portfolios on size and book-to-market (as Fama and French (1993) originally did), the simple three-factor model achieves the best subjective MSPE (around half of the subjective MSPE obtained with the largest models). With 2,875 (5×5) portfolios, however, the ranking reverses, and larger models now have the lowest subjective MSPE. (Appendix C.2 presents several robustness checks with different models, sample sizes, and subsets of portfolios.)

[19] We follow Feng et al. (2020) and use monthly returns from July 1976 to December 2017.

[20] In the standard Fama-French two-pass regressions, these factor loadings are estimated from asset-by-asset time series regressions of excess returns on factors (Bai and Zhou, 2015). For simplicity, we ignore the estimation error and treat the estimated factor loadings as the true factor loadings. This allows us to stay within our simple linear regression framework. The competing models are the different subsets of the factors that different agents believe to be relevant to explain the cross-section of asset expected returns.

[21] There is some discussion in the literature about what is the right unit of observation to test asset-pricing models: see the discussion in Ang et al. (2020).

[22] A 5×5 bivariate portfolio refers to the common practice of grouping stocks by the quintiles of the cross-sectional distribution of both size and book-to-market.

[23] The sorting is based on size and each of the 115 factors marked with an asterisk in the table in Appendix C.3. The dataset was obtained directly from the replication files provided by Feng et al. (2020).

[24] These are the investment and profitability factors of Hou et al. (2015), the "robust minus weak" factor of Fama and French (2015), the intermediary risk factor of He et al. (2017), and the "Quality Minus Junk"

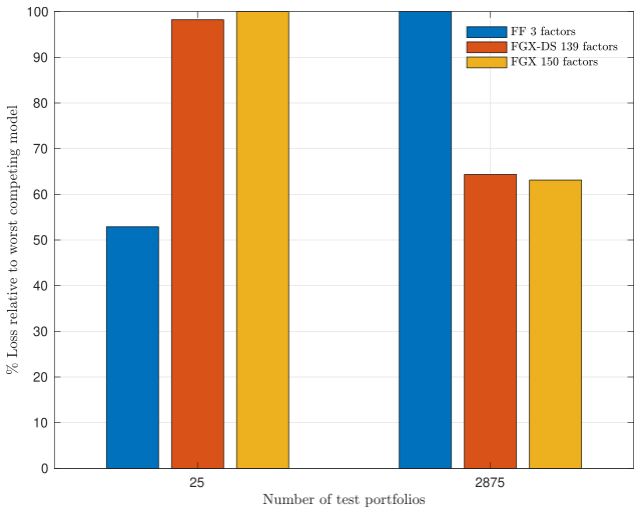

Figure 3: Competing (Factor) Models

Our implications continue to hold even if each of the asset-pricing models that the agents are considering are incorrectly specified. The details are explained below.

First, our theoretical results for small n do not impose any restriction on the true DGP. Consequently, when n is small, our results predict that models based on only a few pricing factors will be selected regardless of the true dependence (or lack thereof) of expected returns on factor exposures. This is confirmed in our application, where the simple Fama-French models with three or five factors indeed outperform more complicated models (in terms of subjective MSPE), provided the only available data are the average returns of the 25 Fama-French test portfolios and their factor exposures (see Figure 3 above and Figure 4 in Appendix C.2).

Second, our theoretical results for large n do impose some restrictions on the true DGP. However, as we explain in Section 5, these restrictions are mainly intended to guarantee that the maximum likelihood estimator of (β, σ^2) based on the agents' linear regression model satisfies some regularity conditions (e.g., asymptotic normality). In the context of our application, these regularity conditions are likely to hold even if there is no dependence between expected returns and factor exposures. Our results predict that the winning model when n is large will have more pricing factors relative to the population's best linear predictor of expected returns given the factor exposures (and such a best linear predictor may have all of its coefficients equal to zero). Our empirical results are in line with our predictions: the model that uses the whole library of 150 pricing factors outperforms other popular models, when the data available are the expected returns of the 2,875 (5×5) bivariate sorted portfolios in Feng et al. (2020).

factor of Asness et al. (2019).

A General Model of Scientific Progress. This discussion suggests a simple model of scientific progress. There is public interest in predicting a variable y (e.g., asset returns averaged over time, as in the example above) as a function of observed covariates x (the various factor exposures that may influence assets' returns). There are competing scientists, each described by a prior belief about the world; these priors differ in what covariates they think are relevant for predicting y. Data (publicly) accumulates as i.i.d. draws from the unknown DGP.

Suppose each model's success also depends on its subjective MSPE. Scientists who believe their model has a low prediction error would be more forceful about it, staking their career on its predictions. Others whose subjective MSPE is high may be worried about mistakes and damage to their reputation. Practitioners or politicians, who may cite scientific research to justify their actions, may be more prone to adopt models with low subjective MSPE. In the context of the example above, we can think of different researchers trying to convince scholars working on the asset-pricing literature that a collection of factors ought to be viewed as the "state-of-the-art." Researchers whose models have better MSPE may be able to make a stronger case for their findings, and publish their work in higher-impact outlets. Model competition in this case (e.g., what factor models get published) selects the models with lower MSPE.

With these assumptions, our results suggest the following dynamic of scientific progress. In the early stages of a field, when data is relatively scarce, overly simple models prevail—including frameworks that exclude relevant covariates. Over time, more and more data accumulates, and more nuanced models come into vogue, involving ever-increasing collections of covariates. Overly-simple models are then discarded, since they are unable to fit the data as well as larger ones; the "scientific paradigm," understood as the collection of relevant variables, becomes more complex. This is in line with casual observation and with dynamics described in epistemology. For example, this aligns with what Kuhn (1962) describes as the path of progress of "normal" science, i.e., after a dominant paradigm has been established.[25] These dynamics are precisely those we documented above in the case of competing factor models.

One final remark concerns the relevance of our large n results in applications, since, in this case, the difference in predictions between overcomplicated and correct models converges to zero. We argue that the distortions resulting from overcomplicated models winning the competition remain relevant. To illustrate, consider our asset pricing application and our

[25] Changing paradigms are outside the scope of this work—see, e.g., Ortoleva (2012) for a model of a non-Bayesian decision-maker who changes paradigm (selects a new prior) upon receiving information that is unexpected according to their current prior.

discussion regarding our simple model of scientific progress.

Suppose that the true model contains only one factor (e.g., the original Capital Asset Pricing Model) and that the only other competing model uses all of the 150 factors in the factor library. Our results imply (see the discussion after Equation 9) that if the risk premium for the single factor in the true model is small, then the probability that the higher-dimensional model defeats the true model is approximately $\mathbb{P}(\chi^2_{149} > 149) \approx 51\%$. Therefore, in large samples, the higher-dimensional model wins the competition with probability close to a half.

Although both models will give similar predictions for expected returns, they are qualitatively very different, and suggest very different directions for the progress of the asset pricing literature.

When the model with only one factor wins the competition, there is a significant data reduction. As explained by Cochrane (2011), this means that new theories of asset pricing only have to explain the risk premium to one particular factor, and not the whole cross-sectional difference in expected stock returns. Moreover, looking for new factors and judging whether a new factor adds explanatory power to asset pricing becomes less important since 150 other factors already proved ineffective.

When the model with 150 factors wins, it points to a very different path for asset pricing research. Since, in the winning model, all 150 factors in the library are relevant for prediction, it is reasonable to expect that yet more factors could emerge, and it becomes important—as explained by Feng et al. (2020)—to study the marginal contribution of new factors relative to the vast set of existing ones.

Therefore, even if an overcomplicated model generates similar predictions as the true model, the outcome of the model competition might matter. In our particular application, this seems to explain the current state of the asset pricing literature. The current competition in the literature is not between the Capital Asset Pricing Model and a model using the whole factor library. However, our illustrative example seems to capture the tension between some low-dimensional asset-pricing models—such as the five-factor model of Fama and French (2015)—and higher-dimensional models—such as one based on the 139 factors selected in Giglio and Xiu (2019).

7 Related Literature

A large body of literature has studied model misspecification in individual decision-making, with famous examples like overconfidence and correlation neglect. A few recent theoretical contributions to this enormous literature include Heidhues et al. (2018) and Ortoleva and

Snowberg (2015), to which we refer for further references. In misspecified learning settings, "feedback loops" between the agents' misspecified beliefs and the action they take add further technical challenges—see, e.g., Fudenberg et al. (2017), Fudenberg et al. (2020), and Heidhues et al. (2020).

Recent works have studied the implications of agents with misspecified models in various strategic settings. For instance, Bohren (2016), Bohren and Hauser (2017), Frick et al. (2019b), and Frick et al. (2019a) study social learning when agents have misspecified models that cause them to misinterpret other agents' actions. Mailath and Samuelson (2019) study a stylized prediction market where Bayesian agents have different models (defined as different partitions of a common state space) and discuss the possibility of information aggregation.

Recent works consider specifically the outcomes when agents' models are misspecified in the sense we study here, i.e., there is a payoff-relevant dependent variable, and agents either include irrelevant independent variables or exclude dependent variables. Schwartzstein and Sunderam (2021) consider this in the context of persuasion, where competing persuaders may "overfit" the data to better persuade a receiver. Levy et al. (2019) study a political economy setting where there are both simple and complicated world views, and ask whether political competition disciplines overly simplistic world views, finding instead that they recur in dynamic settings. Finally, He and Libgober (2020) ask whether (and what kind of) misspecifications can be evolutionarily stable. These works find reasons for why misspecified models may survive (overfit models in the first case, simple models in the latter two), although the exact mechanism is different than ours. Recent work also considers the possibility that agents with misspecified models may be able to realize it, and characterize the kinds of misspecifications that survive—see, e.g., Fudenberg and Lanzani (2020), who consider an evolutionary framework, or Gagnon-Bartsch et al. (2021), who study a setting where the agent perceives their errors through the framework of their own model.

In strategic settings, Esponda and Pouzo (2016) define a learning-based solution concept ("Berk-Nash Equilibrium") for games in which agents' beliefs are misspecified. More broadly, solution concepts have been posited for settings where agents suffer from some sort of misspecification, including well-known examples like analogy-based equilibrium (Jehiel, 2005) and cursed equilibrium (Eyster and Rabin, 2005).

Several works consider outcomes when some agents behave in a way that can be construed as coming from a misspecified model. For instance, in Spiegler (2006, 2013), society misunderstands the relationship between outcomes and the actions of strategic agents, which affects the actions these agents take in equilibrium and resulting outcomes (studied in the context of a market for quacks or its implications for political reforms). Levy et al. (2019) study a dynamic model of political competition where agents have different (misspecified)

models of the world; the study uses this model to provide a foundation for the recurrence of populism. Liang (2018) studies outcomes in games of incomplete information where agents behave like statisticians and have limited information.[26]

A novel approach to modeling misspecification in economic theory is the directed acyclic graph approach; see Pearl (2009). This is exploited in a single-person decision framework in Spiegler (2016), which studies a single decision maker with a misspecified causal model and large amounts of data. The paper shows that the decision maker may evaluate actions differently than their long-run frequencies, and exhibit artifacts such as "reverse causation" and coarse decision making. This approach is then used in Eliaz and Spiegler (2018) who propose a model of competing narratives. A narrative is a causal model that maps actions into consequences, including other random, unrelated variables. An equilibrium notion is defined, and the paper studies the distribution of narratives that is obtained in equilibrium.

Finally, the understanding that agents should be cognizant that their models may be misspecified has also led to new approaches in mechanism design, where the designer accounts for misspecification in various ways. The literature on robust mechanism design (beginning with the seminal Bergemann and Morris 2005) provides foundations for using stronger solution concepts. Madarász and Prat (2017) show that an optimal mechanism may perform very poorly if the planner's model is even slightly misspecified, and they identify a class of near optimal mechanisms that degrade gracefully. Works such as Chassang (2013) and Carroll (2015) develop optimal "robust" contracts and contrast to classical optimal contracting.

Since one natural application of our model is an auction, our results are related to Atakan and Ekmekci (2014), who consider the competitive sale of assets whose value depends on how they are utilized.[27] The successful bidder chooses an action that determines, together with the state of the world, the payoff generated by the asset. They focus on a setting where bidders have a common prior but observe private signals. Their main result is the possibility of (complete) failure of information aggregation. Our model is similar in that the value of the object depends on an action taken by the agent. However, our work considers a complementary environment where all bidders observe the same information but have different priors. Information aggregation is ruled out by assumption, and our key theme is model selection.

We assume that agents have different priors and are fully aware they have different priors: that is to say, our agents agree to disagree. This assumption has been used in economic theory at least since Harrison and Kreps (1978). We refer the reader to Morris (1995) for a discussion

[26] There is a larger literature that studies the outcomes when agents are modeled as statisticians or machine learners, e.g., Al-Najjar (2009), Al-Najjar and Pai (2014), Acemoglu et al. (2016), and Cherry and Salant (2018).

[27] Bond and Eraslan (2010) study a trading environment with a similar feature.

of the common and heterogeneous prior traditions in economic theory. Heterogenous priors have been used in a number of applications in bargaining (Yildiz, 2003), trade (Morris, 1994), financial markets (Scheinkman and Xiong, 2003; Ottaviani and Sørensen, 2015), and more.

Relation to Model Selection. Large bodies of literature in statistics, econometrics, and machine learning study model selection methods and provide normative foundations; see Claeskens and Hjort (2008) and Burnham and Anderson (2003) for textbook overviews. Popular approaches include, for example, the C_p criterion of Mallows (1973), the Akaike Information Criterion (AIC) of Akaike (1974), and the Bayes Information Criterion (BIC) of Schwarz (1978). We have demonstrated the connection between our large data results and the AIC introduced in Akaike (1974), in particular, to the asymptotic properties of the AIC characterized in the seminal paper of Nishii (1984).

While some of our asymptotic results are reminiscent of the model selection literature, there are three important differences. First, the aims of this literature are very different from ours. Ours is a positive approach of studying which model emerges from a competition between Bayesian agents with misspecified models. The approach in the model selection literature is instead *normative*: various methods of model selection are proposed and studied with a view to avoiding over-fitting and/or selecting "good" models according to some metric. The results we are aware of speak to the asymptotic efficiency of these techniques. Second, not only are our results derived from a completely different setting, but they are also proven with different techniques. Third, the connection is limited to the large-data result. We are not aware of any analogs to our small-sample results.

8 Discussion and Conclusion

A variable of interest is related to a vector of covariates. Different agents have different models of this relationship: in particular they rule in/rule out different covariates as being potentially related to prediction. All agents observe a common dataset of size n, drawn from the true DGP. We ask: Who is the agent with the highest confidence in their own predictive ability, in the form of the lowest mean squared prediction error according to their own subjective posterior? We study the relationship between sample size and the dimension of the winning model. This applies to all cases in which confidence in predictive ability affects selection. We show results of two kinds.

First, when n is small, models that employ few covariates may take the lead, even if the true DGP is more complex. To establish this result formally, we use Normal-Inverse-Gamma-Inverse-Wishart priors. Second, when n is large, misspecified models (i.e., models that rule

out an observable that is relevant for prediction) never win. However, high-dimensional models that include irrelevant covariates (but do not exclude relevant ones) may continue to win. Our results show that the effect of the prior on model competition does not vanish in large samples. These results hold for a very general class of priors and true DGPs.

Finally, we give two applications. First, we apply our results to a model of entry: entrepreneurs decide whether to enter a new market, households decide whether to invest in a new asset class. We show that, insofar as prediction error of future variables is relevant for profitability, our results suggest a new margin of selection: when data is relatively scarce, agents with simpler models will be overrepresented in the entry decision. Our second application is to understand the proliferation of factors that explain the cross-sectional variation of expected stock returns in the asset-pricing literature. We show how the increase in the number of test portfolios used to compute the cross-sectional regressions mechanically favors models with several factors.

References

Acemoglu, D., V. Chernozhukov, and M. Yildiz (2016): "Fragility of asymptotic agreement under Bayesian learning," *Theoretical Economics*, 11, 187–225.

Akaike, H. (1974): "A new look at the statistical model identification," *IEEE transactions on automatic control*, 19, 716–723.

Al-Najjar, N. I. (2009): "Decision makers as statisticians: Diversity, ambiguity, and learning," *Econometrica*, 77, 1371–1401.

Al-Najjar, N. I. and M. M. Pai (2014): "Coarse decision making and overfitting," *Journal of Economic Theory*, 150, 467–486.

Alonso, R., W. Dessein, and N. Matouschek (2008): "When does coordination require centralization?" *American Economic Review*, 98, 145–79.

Ang, A., J. Liu, and K. Schwarz (2020): "Using Stocks or Portfolios in Tests of Factor Models," *Journal of Financial and Quantitative Analysis*, 55, 709–750.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2019): "Quality minus junk," *Review of Accounting Studies*, 24, 34–112.

Atakan, A. E. and M. Ekmekci (2014): "Auctions, actions, and the failure of information aggregation," *American Economic Review*, 104.

Bai, J. and G. Zhou (2015): "Fama–MacBeth two-pass regressions: Improving risk premia estimates," *Finance Research Letters*, 15, 31–40.

Barberis, N., C. Harvey, and N. Shephard (2015): "Overfitting and its impact on the Investor," *Man AHL Academic Advisory Board*.

Bergemann, D. and S. Morris (2005): "Robust mechanism design," *Econometrica*, 73, 1771–1813.

Berk, R. H. (1970): "Consistency a posteriori," *The Annals of Mathematical Statistics*, 894–906.

Bishop, C. M. (2006): *Pattern recognition and machine learning*, springer.

Bohren, J. A. (2016): "Informational herding with model misspecification," *Journal of Economic Theory*, 163, 222–247.

Bohren, J. A. and D. Hauser (2017): "Bounded rationality and learning: A framework and a robustness result," *Working Paper, University of Pennsylvania*.

Bond, P. and H. Eraslan (2010): "Information-based trade," *Journal of Economic Theory*, 145, 1675–1703.

Burnham, K. P. and D. R. Anderson (2003): *Model selection and multimodel inference: a practical information-theoretic approach*, Springer Science & Business Media.

Carroll, G. (2015): "Robustness and linear contracts," *American Economic Review*, 105, 536–63.

Chassang, S. (2013): "Calibrated incentive contracts," *Econometrica*, 81, 1935–1971.

Chen, Z., C. Li, and W. Sun (2020): "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," *Journal of Computational and Applied Mathematics*, 365, 112395.

Cherry, J. and Y. Salant (2018): "Statistical Inference in Games," Tech. rep., mimeo.

Claeskens, G. and N. Hjort (2008): "Model selection and model averaging," *Cambridge Books*.

Cochrane, J. H. (2011): "Presidential address: Discount rates," *The Journal of finance*, 66, 1047–1108.

Cooper, A. C., C. Y. Woo, and W. C. Dunkelberg (1988): "Entrepreneurs' perceived chances for success," *Journal of Business Venturing*, 3, 97–108.

Eliaz, K. and R. Spiegler (2018): "A Model of Competing Narratives," *CEPR Discussion Paper No. DP13319*.

Esponda, I. and D. Pouzo (2016): "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 84, 1093–1130.

Eyster, E. and M. Rabin (2005): "Cursed equilibrium," *Econometrica*, 73, 1623–1672.

Fama, E. F. and K. R. French (1993): "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 33, 3–56.

——— (2015): "A five-factor asset pricing model," *Journal of financial economics*, 116, 1–22.

Fama, E. F. and J. D. MacBeth (1973): "Risk, return, and equilibrium: Empirical tests," *Journal of political economy*, 81, 607–636.

Feng, G., S. Giglio, and D. Xiu (2020): "Taming the factor zoo: A test of new factors," *The Journal of Finance*, 75, 1327–1370.

Frick, M., R. Iijima, and Y. Ishii (2019a): "Dispersed Behavior and Perceptions in Assortative Societies," .

——— (2019b): "Misinterpreting Others and the Fragility of Social Learning," *Cowles Foundation Discussion Paper*.

Fudenberg, D. and G. Lanzani (2020): "Which misperceptions persist?" *Available at SSRN*.

Fudenberg, D., G. Lanzani, and P. Strack (2020): "Limits Points of Endogenous Misspecified Learning," *Available at SSRN*.

Fudenberg, D., G. Romanyuk, and P. Strack (2017): "Active learning with a misspecified prior," *Theoretical Economics*, 12, 1155–1189.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2021): "Channeled Attention and Stable Errors," *Working Paper*.

Giglio, S. and D. Xiu (2019): "Asset pricing with omitted factors," *Chicago Booth Research Paper*.

Greene, W. H. (2018): *Econometric Analysis*, vol. 8th Edition, Pearson.

Hansen, B. (2021): "Econometrics," *A textbook draft available online at www. ssc. wisc. edu/˜ bhansen/econometrics/Econometrics. pdf*.

Harrison, J. M. and D. M. Kreps (1978): "Speculative investor behavior in a stock market with heterogeneous expectations," *The Quarterly Journal of Economics*, 92, 323–336.

He, K. and J. Libgober (2020): "Evolutionarily Stable (Mis) specifications: Theory and Applications," *arXiv preprint arXiv:2012.15007*.

He, Z., B. Kelly, and A. Manela (2017): "Intermediary asset pricing: New evidence from many asset classes," *Journal of Financial Economics*, 126, 1–35.

Heidhues, P., B. Kőszegi, and P. Strack (2018): "Unrealistic expectations and misguided learning," *Econometrica*, 86, 1159–1214.

Heidhues, P., B. Koszegi, and P. Strack (2020): "Convergence in Models of Misspecified Learning," .

Hogg, R. V., J. W. Mckean, and C. D. Allen (2006): *Introduction To Mathematical Statistics*, Pearson Education India.

Hou, K., C. Xue, and L. Zhang (2015): "Digesting anomalies: An investment approach," *The Review of Financial Studies*, 28, 650–705.

Jehiel, P. (2005): "Analogy-based expectation equilibrium," *Journal of Economic theory*, 123, 81–104.

Jensen, M. C., F. Black, and M. S. Scholes (1972): "The capital asset pricing model: Some empirical tests," .

Kass, R., L. Tierney, and J. B. Kadane (1990): "The validity of posterior expansions based on Laplaces method," in *Bayesian and Likelihood Methods in Statistics and Econometrics*, ed. by S. Geisser, J. Hodges, S. Press, and A. Zellner, vol. 7, 473.

Kim, K. T. and S. D. Hanna (2021): "Investment literacy, overconfidence and cryptocurrency investment," *Working Paper*.

Kleijn, B. and A. Van der Vaart (2012): "The Bernstein-von-Mises theorem under misspecification," *Electronic Journal of Statistics*, 6, 354–381.

Koellinger, P., M. Minniti, and C. Schade (2007): ""I think I can, I think I can": Overconfidence and entrepreneurial behavior," *Journal of economic psychology*, 28, 502–527.

Kozak, S., S. Nagel, and S. Santosh (2018): "Interpreting Factor Models," *Journal of Finance*, 73, 1183–1223.

Kuhn, T. S. (1962): *The structure of scientific revolutions*, University of Chicago press.

Levy, G., R. Razin, and A. Young (2019): "Misspecified Politics and the Recurrence of Populism," Tech. rep., Working Paper.

Liang, A. (2018): "Games of Incomplete Information Played by Statisticians," *Working paper, University of Pennsylvania*.

Madarász, K. and A. Prat (2017): "Sellers with misspecified models," *The Review of Economic Studies*, 84, 790–815.

Mailath, G. J. and L. Samuelson (2019): "The Wisdom of a Confused Crowd: Model-Based Inference," *Cowles Foundation Discussion Paper*.

Mallows, C. L. (1973): "Some Comments on C_P," *Technometrics*, 15, 661–675.

Marschak, J. and R. Radner (1972): *Economic Theory of Teams.*

McCullagh, P. (2002): "What is a statistical model?" *The Annals of Statistics*, 30, 1225–1310.

Milgrom, P. and J. Roberts (1992): *Economics, organization and management*, Prentice-Hall Englewood Cliffs, NJ.

Morris, S. (1994): "Trade with heterogeneous prior beliefs and asymmetric information," *Econometrica: Journal of the Econometric Society*, 1327–1347.

——— (1995): "The common prior assumption in economic theory," *Economics & Philosophy*, 11, 227–253.

Munim, Z. H., M. H. Shakil, and I. Alon (2019): "Next-Day Bitcoin Price Forecast," *Journal of Risk and Financial Management*, 12.

Nishii, R. (1984): "Asymptotic properties of criteria for selection of variables in multiple regression," *The Annals of Statistics*, 758–765.

Odean, T. (1999): "Do investors trade too much?" *American economic review*, 89, 1279–1298.

Ortoleva, P. (2012): "Modeling the change of paradigm: Non-Bayesian reactions to unexpected news," *American Economic Review*, 102, 2410–36.

Ortoleva, P. and E. Snowberg (2015): "Overconfidence in political behavior," *American Economic Review*, 105, 504–35.

Ottaviani, M. and P. N. Sørensen (2015): "Price reaction to information with heterogeneous beliefs and wealth effects: Underreaction, momentum, and reversal," *American Economic Review*, 105, 1–34.

Pearl, J. (2009): *Causality*, Cambridge university press.

Scheinkman, J. A. and W. Xiong (2003): "Overconfidence and speculative bubbles," *Journal of political Economy*, 111, 1183–1220.

Schwartzstein, J. and A. Sunderam (2021): "Using models to persuade," *American Economic Review*, 111, 276–323.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Ann. Statist.*, 6, 461–464.

Spiegler, R. (2006): "The market for quacks," *The Review of Economic Studies*, 73, 1113–1131.

——— (2013): "Placebo reforms," *American Economic Review*, 103, 1490–1506.

——— (2016): "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 131, 1243–1290.

Statman, M., S. Thorley, and K. Vorkink (2006): "Investor overconfidence and trading volume," *The Review of Financial Studies*, 19, 1531–1565.

Yildiz, M. (2003): "Bargaining without a common prior—an immediate agreement theorem," *Econometrica*, 71, 793–811.

A Main Appendix

A.1 Second-price auction

Consider a second-price auction, where, like in Atakan and Ekmekci (2014), the winner of the auction gets to choose an action that affects the value of the asset. Specifically, the action has a value that depends on her ability to predict a given variable, as in the examples given in the introduction. Formally, fixing the environment defined above (DGP, agents, etc.), consider a game with the following timing:

1. Nature draws $\theta_0 \in \Theta$;
2. All agents see a common dataset D_n drawn according to Q_{θ_0};
3. Agents submit bid in a sealed-bid second-price auction;
4. The winner observes x randomly drawn according to P and chooses a real-valued action a;
5. The winner gets a lump sum payoff of $M-(y-a)^2$, where M is a large positive number.

Every bidder seeks to minimize the expected value $M-(y-a)^2$, leading to the expected loss function discussed above.

Because agents see a common data set, an agent with prior π has an expected value of $M - L^*(\pi, D_n)$ for winning. In the standard dominant equilibrium, the winning agent is the one with the highest value: since M is common across agents, the winner is thus the agent with the lowest expected loss (according to her own prior) given the observed data. Notice that since all agents observe the same dataset, and thus there is no asymmetric information (only heterogenous priors), winner's-curse-type considerations do not apply.[28]

A.2 Proof of Lemma 1

Lemma 1. *Suppose that $\beta_{J(\pi)}$ is independent of P under the posterior distribution. The agent's subjective MSPE can be decomposed as:*

$$L^*(\pi, D_n) = \underbrace{\mathbb{E}_\pi\left[\sigma^2 | D_n\right]}_{\textit{Model Fit}} + \underbrace{\operatorname{tr}\left(\mathbb{V}_\pi[\beta_{J(\pi)}|D_n]\ \mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}] \mid D_n\right]\right)}_{\textit{Model Estimation Uncertainty}}, \tag{5}$$

[28] Our results possibly shed light on political competition/ board meetings. While we do not develop these formally, intuitively, these would correspond to an analogous all-pay auction. Agents have different models of how to forecast payoff-relevant unknowns from covariates. The action taken (by the government body or company) depends on this forecast. Agents' willingness to lobby for their model depends on how confident they are in their model, and the amount of effort they spend lobbying influences selection.

where $\mathbb{V}(\cdot)$ *is the variance-covariance operator, and* tr *is the trace operator.*

Proof. Fix a data set D_n. We need to analyze

$$\mathbb{E}_\pi\left[\mathbb{E}_P\left[(x'\beta - f^*_{(\pi,D_n)}(x))^2\right]\middle| D_n\right].$$

From (4), $f^*_{(\pi,D_n)}(x) = \mathbb{E}_\pi[\beta|D_n]'x$. Thus, the expectation above becomes

$$=\mathbb{E}_\pi\left[\mathbb{E}_P\left[(x'(\beta - \mathbb{E}_\pi[\beta|D_n]))^2\right]\middle| D_n\right].$$

Recalling that for a scalar a, $a = \text{tr}(a)$, we have

$$=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\text{tr}[(x'(\beta - \mathbb{E}_\pi[\beta|D_n]))^2]\right]\middle| D_n\right].$$

Then by symmetry and linearity of the trace operator, we can conclude,

$$\begin{aligned}
&=\mathbb{E}_\pi\left[\mathbb{E}_P\left[\text{tr}[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'xx']\right]\middle| D_n\right],\\
&=\mathbb{E}_\pi\left[\text{tr}[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'\mathbb{E}_P[xx']]\middle| D_n\right],\\
&=\text{tr}\left[\mathbb{E}_\pi\left[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'\mathbb{E}_P[xx']\middle| D_n\right]\right].
\end{aligned}$$

Since β and P are independent under the posterior by assumption, we have that,

$$\mathbb{E}_\pi\left[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'\mathbb{E}_P[xx']\middle| D_n\right]$$

equals

$$\mathbb{E}_\pi\left[(\beta - \mathbb{E}_\pi[\beta|D_n])(\beta - \mathbb{E}_\pi[\beta|D_n])'\middle| D_n\right]\mathbb{E}_\pi\left[\mathbb{E}_P[xx']\middle| D_n\right].$$

Finally, by the definition of variance, we have the desired form

$$=\text{tr}(\mathbb{V}_\pi(\beta|D_n)\mathbb{E}_\pi\left[E_P[xx'] \mid D_n\right]).$$ □

A.3 Proof of Lemma 2

Before we delve into the proof, we recall a few facts about the Normal-Inverse-Gamma-Inverse-Wishart distribution which may be useful. In particular, straightforward algebra

shows that given a dataset D_n:

$$\mathbb{E}_\pi[\sigma^2|D_n] = \frac{\frac{2b_0}{n} + \frac{1}{n}(Y'Y - Y'X_J(X_J'X_J + \gamma_0|J|\mathbb{I}_J)^{-1}X_J'Y)}{\frac{2a_0}{n} + 1 - \frac{2}{n}}, \tag{10}$$

$$\mathbb{V}_\pi[\beta_J|D_n] = \mathbb{E}_\pi\left[\sigma^2|D_n\right](X_J'X_J + (\gamma_0|J|)\mathbb{I}_{|J|})^{-1}. \tag{11}$$

The posterior distribution of $\Sigma_{J(\pi)}$ after observing the data D_n is given as:

$$\Sigma_{J(\pi)}|D_n \sim \text{Inverse-Wishart}(X_{J(\pi)}'X_{J(\pi)} + \gamma_0|J(\pi)|\mathbb{I}_{J(\pi)}, n + 2|J(\pi)| + 1).$$

Therefore,

$$\mathbb{E}_\pi[\Sigma_{J(\pi)}|D_n] = \frac{X_{J(\pi)}'X_{J(\pi)} + \gamma_0|J(\pi)|\mathbb{I}_{J(\pi)}}{n + |J(\pi)|}. \tag{12}$$

Lemma 2. *Consider a prior π such that the implied distribution over $(\beta_J, \sigma^2, \Sigma_J)$ are as in Definition 1. Then*

$$L^*(\pi, D_n) = \underbrace{\mathbb{E}_\pi[\sigma^2|D_n]}_{\textit{Model Fit}} + \underbrace{\mathbb{E}_\pi[\sigma^2|D_n]\left(\frac{|J(\pi)|}{n + |J(\pi)|}\right)}_{\textit{Model Estimation Uncertainty}}. \tag{6}$$

Proof. Note that Remark 1 implies that the condition of Lemma 1 is satisfied and therefore we have the decomposition, i.e., (5):

$$L^*(\pi, D_n) = \mathbb{E}_\pi[\sigma^2|D_n] + \text{tr}\left(\mathbb{V}_\pi(\beta|D_n)\mathbb{E}_\pi[\Sigma|D_n]\right).$$

Note that for any model $J(\pi)$, the mean of $\Sigma_{J(\pi)}$ under the prior is $\gamma_0\mathbb{I}_{|J(\pi)|}$, and therefore the prior loss is

$$\mathbb{E}_\pi[\sigma^2]\left(1 + \text{tr}(\mathbb{I}_{|J(\pi)|})/|J(\pi)|\right) = 2\mathbb{E}_\pi[\sigma^2].$$

This means all models have the same ex-ante loss prior to observing the data. Substituting in $V_\pi(\beta|D_n)$ from (11), and $\mathbb{E}_\pi[\Sigma_{J(\pi)}|D_n]$ from (12) into (5) we have that

$$L^*(\pi, D_n) = \mathbb{E}_\pi[\sigma^2|D_n]\left(1 + \frac{|J(\pi)|}{n + |J(\pi)|}\right),$$

as desired. □

A.4 Proof of Proposition 1

Proposition 1. *Suppose all priors are as in Definition 1 with shared hyperparameters. Suppose also that for every single covariate model, i.e., every J such that $|J| = 1$, there is an agent with that model, and that all agents use at least one covariate. If $n = 1$, then the agent with lowest subject MSPE is an agent with a single covariate model.*

Proof. Denote the single datapoint as $D_1 = (Y, X)$, where $Y \in \mathbb{R}$ and $X \in \mathbb{R}^{1\times k}$ (k is the number of covariates), where $X = (x_1, \ldots, x_k)$. First, observe that for any agent j with a single explanatory variable κ in his model (denoted x_κ),

$$\begin{aligned} L^*(\pi_j, D_1) &= \frac{b_0 + \frac{1}{2}\left(y^2 - \frac{y^2 x_\kappa^2}{x_\kappa^2 + \gamma_0}\right)}{a_0 - \frac{1}{2}} \left(1 + \frac{1}{2}\right), \\ &= \frac{b_0 + \frac{1}{2}\frac{y^2\gamma_0}{x_\kappa^2 + \gamma_0}}{a_0 - \frac{1}{2}} \left(1 + \frac{1}{2}\right). \end{aligned}$$

The winning agent among the single variable models will therefore clearly be the agent with the variable κ that maximizes x_κ^2. Without loss of generality, call this variable 1.

To economize on notation, now consider the full model with all the explanatory variables, it will be clear from the logic that this argument will work for any model larger than a single variable. For an agent j with all k variables, we know that

$$L^*(\pi_j, D_1) = \frac{b_0 + \frac{y^2}{2}\left(1 - X(X'X + \gamma_0 k \mathbb{I}_k)^{-1} X'\right)}{a_0 - \frac{1}{2}} \left(1 + \frac{k}{1+k}\right).$$

We show that this model always loses to the "best" single variable model. To do this, note first that for any $k > 1$

$$\frac{k}{1+k} > 1/2.$$

Consequently, it is sufficient to show:

$$\left(1 - X(X'X + \gamma_0 k \mathbb{I}_k)^{-1} X'\right) \geq \frac{\gamma_0}{x_1^2 + \gamma_0}.$$

In what follows, we assume that $X \neq \mathbf{0}$ (as the inequality clearly holds when X is a vector of zeros). Algebra shows that

$$\begin{aligned} &\left(1 - X(X'X + \gamma_0 k \mathbb{I}_k)^{-1} X'\right) \geq \frac{\gamma_0}{x_1^2 + \gamma_0}, \\ \iff &X(X'X + \gamma_0 k \mathbb{I}_k)^{-1} X' \leq \frac{x_1^2}{x_1^2 + \gamma_0}. \end{aligned}$$

Observe that $X(X'X + \gamma_0 k\mathbb{I}_k)^{-1}X'$ is a scalar. We know that for a scalar, $a = \text{tr}(a)$. Therefore:

$$\begin{aligned}
&X(X'X + \gamma_0 k\mathbb{I}_k)^{-1}X', \\
=&\text{tr}[X(X'X + \gamma_0 k\mathbb{I}_k)^{-1}X'], \\
=&\text{tr}[(X'X + \gamma_0 k\mathbb{I}_k)^{-1}X'X], \\
=&\text{tr}\left[(\frac{1}{\gamma_0 k}X'X + \mathbb{I}_k)^{-1}\frac{1}{\gamma_0 k}X'X\right].
\end{aligned}$$

Denote $\frac{1}{\gamma_0 k}X'X$ as A. Substituting

$$=\text{tr}[(A + \mathbb{I}_k)^{-1}A].$$

Observe that if λ is an eigenvalue of A, then $\frac{\lambda}{1+\lambda}$ is an eigenvalue of $(A + \mathbb{I}_k)^{-1}A$. To see this, suppose v is an eigenvector of A with eigenvalue λ. Then,

$$\begin{aligned}
&Av = \lambda v, && (13)\\
\implies &(A + \mathbb{I}_k)v = (\lambda + 1)v, \\
\implies &(A + \mathbb{I}_k)^{-1}v = \frac{1}{1+\lambda}v. && (14)
\end{aligned}$$

Therefore we have that

$$\begin{aligned}
(A + \mathbb{I}_k)^{-1}Av &= (A + \mathbb{I}_k)^{-1}(\lambda v), && \text{(by (13))} \\
&= \frac{\lambda}{1+\lambda}v. && \text{(by (14))}
\end{aligned}$$

Substituting this in, we have

$$\text{tr}[(A + \mathbb{I}_k)^{-1}A] = \sum_{i=1}^{k}\frac{\lambda_i}{1+\lambda_i}.$$

Therefore we are left to show that

$$\sum_{i=1}^{k}\frac{\lambda_i}{1+\lambda_i} \leq \frac{x_1^2}{x_1^2 + \gamma_0}.$$

Here λ_i's are the eigenvalues of $\frac{1}{\gamma_0 k}X'X$. This implies that $\sum_i \lambda_i = \frac{1}{\gamma_0 k}\sum_i x_i^2$.

Note that $X'X$ is not full rank, indeed, its null space is of dimension $k - 1$. Therefore

it has $k-1$ multiplicity eigenvalue of 0. The unique non-zero eigenvalue must then be $\frac{1}{\gamma_0 k}\sum_i x_i^2$.

Substituting in, we have

$$\begin{aligned}\sum_{i=1}^{k}\frac{\lambda_i}{1+\lambda_i} =& \frac{\frac{1}{\gamma_0 k}\sum_i x_i^2}{\frac{1}{\gamma_0 k}\sum_i x_i^2+1},\\ =& \frac{\frac{1}{k}\sum_i x_i^2}{\frac{1}{k}\sum_i x_i^2+\gamma_0},\\ \leq& \frac{x_1^2}{x_1^2+\gamma_0}.\end{aligned}$$

where the last inequality follows since we assumed that $x_1^2 = \max_i\{x_i^2 : 1\leq i\leq k\}$. □

A.5 Proof of Proposition 2

Proposition 2. *Let Π be a finite set of agents' priors that satisfy Definition 1 with shared hyperparameters (a_0, b_0, γ_0). Let $|\underline{J}|$ be the size of the smallest model in this set. Fix the size of the dataset n. For any $p\in(0,1)$, there exists b_0 large enough so*

$$\mathbb{P}\left(D_n : \exists \pi^* \in \operatorname*{argmin}_{\pi\in\Pi} L^*(\pi, D_n) \ s.t.\ |J(\pi^*)| = |\underline{J}|\right) > p,$$

i.e., with probability at least p over datasets D_n, the agent with the lowest subjective MSPE has the smallest size model among all the agents.

Proof. Fix any prior $\pi\in\Pi$ such that $|J(\pi)| = |\underline{J}|$. Let $\Pi'\subset\Pi$ be the set of priors with size larger than $|\underline{J}|$, i.e. $\Pi' = \{\pi' : \pi'\in\Pi \text{ and } |J(\pi')| > |\underline{J}|\}$.

From (7), we have that for any other prior $\pi'\in\Pi'$:

$$\begin{aligned}& L^*(\pi, D_n) \geq L^*(\pi', D_n),\\ \iff& \frac{\mathbb{E}_\pi[\sigma^2|D_n]}{\mathbb{E}_{\pi'}[\sigma^2|D_n]} \geq \frac{\left(1+\frac{|J(\pi')|}{n+|J(\pi')|}\right)}{\left(1+\frac{|J(\pi)|}{n+|J(\pi)|}\right)},\end{aligned}$$

We know from (10) that the left hand side,

$$\frac{\mathbb{E}_\pi[\sigma^2|D_n]}{\mathbb{E}_{\pi'}[\sigma^2|D_n]} = \frac{2b_0 + (Y'Y - Y'X_{J(\pi)}(X'_{J(\pi)}X_{J(\pi)} + \gamma_0|J|\mathbb{I}_J)^{-1}X'_{J(\pi)}Y)}{2b_0 + (Y'Y - Y'X_{J(\pi')}(X'_{J(\pi')}X_{J(\pi')} + \gamma_0|J|\mathbb{I}_J)^{-1}X'_{J(\pi')}Y)}$$

Therefore as b_0 grows large, we have that (the left hand side) $\frac{\mathbb{E}_\pi[\sigma^2|D_n]}{\mathbb{E}_{\pi'}[\sigma^2|D_n]} \to_P 1$. However, the

right hand side is a constant that is larger than 1 by observation. Therefore, for any given p', there exists b_0 large enough so that $\mathbb{P}(D_n : L^*(\pi, D_n) \leq L^*(\pi', D_n)) > p'$.

Suppose we fix b_0 large enough so that $\forall \pi' \in \Pi' : \mathbb{P}(D_n : L^*(\pi, D_n) \leq L^*(\pi', D_n)) > p'$. Then the probability $\mathbb{P}\left(D_n : L(\pi, D_n) \leq L(\pi', D_n)\ \forall \pi' \in \Pi'\right) \geq 1 - |\Pi'|(1 - p')$.[29] Selecting $(1 - p')|\Pi'| \leq (1 - p)$ therefore gives us that

$$\mathbb{P}\left(D_n : L(\pi, D_n) \leq L(\pi', D_n)\ \forall \pi' \in \Pi'\right) > p.$$

Since π is s.t. $J(\pi) = |\underline{J}|$ by assumption, we have the desired result.

□

A.6 Proof of Lemma 3

Let the data $D_n = (Y, X)$, $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times k}$ consist of n i.i.d. draws of y and x. Let $\mathbb{P}$ denote the true joint distribution of (y, x).

The density for $Y|X$ corresponding to the Gaussian linear regression model postulated by an agent with variables in $J \subseteq \{1, \ldots, k\}$ is:

$$f(Y|X_J; \beta_J, \sigma^2) := \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(Y - X_J\beta_J)'(Y - X_J\beta_J)\right). \tag{15}$$

Let

$$\widehat{\beta}_J := (X_J'X_J)^{-1}X_J'Y, \quad \widehat{\sigma}_J^2 := (Y - X_J'\widehat{\beta}_J)'(Y - X_J'\widehat{\beta}_J)/n,$$

denote the Maximum Likelihood estimators of (β_J, σ^2) based on (15).

In what follows, let π denote a joint distribution over $(\beta_J, \sigma^2) \in \mathbb{R}^{|J|} \times \mathbb{R}_+$. Let $\pi(\cdot|D_n)$ denote the posterior density of (β_J, σ^2) based on the likelihood (15) and the prior π. Let $\mathbb{E}_\pi[\sigma^2|D_n]$ denote the posterior expectation of σ^2 under the posterior density.

Lemma 3. *Suppose π is a four-times continuously differentiable, strictly positive density function on (β_J, σ^2). Suppose $\mathbb{E}_\pi[\sigma^2|D_n] < \infty$ almost surely, for n large enough. If $X'X/n$*

[29] This follows from Boole's inequality: For any finite collection of sets $A_1, \ldots A_k$:

$$P(A_1 \cap \ldots \cap A_k) = 1 - P(A_1^c \cup \ldots \cup A_k^c) \geq 1 - \sum_{j=1}^{k} P(A_j^c).$$

Therefore, if $P(A_k) \geq p'$ for all k, then

$$P(A_1 \cap \ldots \cap A_k) \geq 1 - k(1 - p').$$

converges in probability to a positive definite matrix, then

$$\mathbb{E}_\pi[\sigma^2|D_n] = \widehat{\sigma}^2_J + \frac{1}{n}\left(2\widehat{\sigma}^4_J\left\{\left(\frac{\partial\pi}{\partial\sigma^2}(\widehat{\theta}_J)\right)\cdot\frac{1}{\pi(\widehat{\theta}_J)}\right\} + \widehat{\sigma}^2_J(|J|+4)\right) + O_{\mathbb{P}}\left(\frac{1}{n^2}\right),$$

where $\widehat{\theta}_J := (\widehat{\beta}'_J, \widehat{\sigma}^2_J)'$.

Proof. The proof has two main steps. First, we introduce some additional notation. Second, we invoke the results of Kass et al. (1990) and apply them to approximate $\mathbb{E}_\pi[\sigma^2|D_n]$.
Step 0 (Additional Notation): Let $\theta = (\beta'_J, \sigma^2)'$ and

$$h_n(\theta) := -\frac{1}{n}\ln f(Y|X_J);\theta).$$

The (i,j) component of the matrix of second derivatives of $h_n(\theta)$ with respect to θ (the Hessian of the scaled negative log-likelihood) will be denoted as $h_{ij}(\cdot)$. We omit the dependence on n, unless confusion arises. The components of the inverse of the Hessian will be written as $h^{ij}(\cdot)$. Finally, $h_{rsj}(\cdot)$ denotes the partial derivative of h_{rs} with respect to the j-th component of θ.

Step 1 (Asymptotic Expansion of $\mathbb{E}_\pi[\sigma^2|D_n]$): Kass et al. (1990) provide asymptotic expansions for posterior moments of a real-valued function of θ in terms of the maximizers of the likelihood used to compute the posterior.

Consider the function

$$g(\theta) = g((\beta'_J, \sigma^2)') = \sigma^2.$$

Theorem 4 in Kass et al. (1990) implies that under the assumptions of our lemma:

$$\begin{aligned}\mathbb{E}_\pi[g(\theta)|D_n] &= g(\widehat{\theta}_J) + \frac{1}{n}\sum_{1\le i,j\le\dim(\theta)}\left(\frac{\partial g}{\partial\theta_i}(\widehat{\theta}_J)\right)h^{ij}(\widehat{\theta}_J)\left\{\left(\frac{\partial\pi}{\partial\theta_j}(\widehat{\theta}_J)\right)\cdot\right.\\ &\qquad\left.\frac{1}{\pi(\widehat{\theta}_J)} - \frac{1}{2}\sum_{1\le r,s\le\dim(\theta)}h^{rs}(\widehat{\theta}_J)h_{rsj}(\widehat{\theta}_J)\right\}\\ &+ \frac{1}{2n}\sum_{1\le i,j\le\dim(\theta)}h^{ij}(\widehat{\theta}_J)\left(\frac{\partial g}{\partial\theta_i\theta_j}(\widehat{\theta}_J)\right)\\ &+ O_{\mathbb{P}}\left(\frac{1}{n^2}\right).\end{aligned}$$

See equation 2.6 in p. 481 of Kass et al. (1990).

Since

$$\frac{\partial g}{\partial \sigma^2}(\theta) = 1,$$

and also

$$\frac{\partial g}{\partial \theta_i}(\theta) = 0,$$

for any $i < |J| + 1$, the expansion above simplifies to

$$\begin{aligned}\mathbb{E}_\pi[\sigma^2|D_n] &= \widehat{\sigma}_J^2 + \frac{1}{n}\sum_{1\le j\le |J|+1} h^{(|J|+1)j}(\widehat{\theta}_J)\left\{\left(\frac{\partial \pi}{\partial \theta_j}(\widehat{\theta})\right)\cdot \right. \\ &\left. \frac{1}{\pi(\widehat{\theta}_J)} - \frac{1}{2}\sum_{1\le r,s\le |J|+1} h^{rs}(\widehat{\theta}_J)h_{rsj}(\widehat{\theta}_J)\right\} \\ &+O_{\mathbb{P}}\left(\frac{1}{n^2}\right).\end{aligned}$$

We now derive explicit formulae for the Hessian matrix, and its inverse elements. The Hessian matrix of $h_n(\theta)$ equals

$$\begin{pmatrix} \frac{1}{n\sigma^2}X_J{}'X_J & \frac{1}{n\sigma^4}X_J{}'(Y - X_J{}'\beta_J) \\ \frac{1}{n\sigma^4}(Y - X_J{}'\beta_J)'X_J & -\frac{1}{2\sigma^4} + \frac{1}{n\sigma^6}(Y - X_J'\beta_J)'(Y - X_J'\beta_J), \end{pmatrix}$$

and the inverse elements of the Hessian evaluated at $\widehat{\theta}$ are:

$$\begin{pmatrix} \widehat{\sigma}_J^2\left(X_J{}'X_J/n\right)^{-1} & \mathbf{0} \\ \mathbf{0} & 2\widehat{\sigma}_J^4, \end{pmatrix}. \tag{16}$$

This further simplifies the expansion to

$$\begin{aligned}\mathbb{E}_\pi[\sigma^2|D_n] &= \widehat{\sigma}_J^2 + \frac{2\widehat{\sigma}_J^4}{n}\left\{\left(\frac{\partial \pi}{\partial \sigma^2}(\widehat{\theta}_J)\right)\cdot \frac{1}{\pi(\widehat{\theta}_J)}\right. \\ &\left. - \frac{1}{2}\sum_{1\le r,s\le |J|+1} h^{rs}(\widehat{\theta})h_{rs(|J|+1)}(\widehat{\theta}_J)\right\} + O_{\mathbb{P}}\left(\frac{1}{n^2}\right).\end{aligned}$$

Algebra shows that the matrix that collects the terms $h_{rs(|J|+1)}(\widehat{\theta}_J)$ (which corresponds to the derivative of the Hessian matrix with respect to σ^2) equals:

$$\begin{pmatrix} -\frac{1}{\widehat{\sigma}_J^4}\left(X_J{}'X_J/n\right) & \mathbf{0} \\ \mathbf{0} & -\frac{2}{\widehat{\sigma}_J^6}, \end{pmatrix}. \tag{17}$$

The term

$$\sum_{1\leq r,s\leq |J|+1} h^{rs}(\widehat{\theta}_J)h_{rs(|J|+1)}(\widehat{\theta}_J),$$

can be written as the sum of all elements of the Hadamard product between the matrices in (16) and (17). This sum, in turn equals:

$$\text{tr}\left(\begin{pmatrix}\widehat{\sigma}_J^2\left(X_J{}'X_J/n\right)^{-1} & \mathbf{0}\\ \mathbf{0} & 2\widehat{\sigma}_J^4,\end{pmatrix}\begin{pmatrix}-\frac{1}{\widehat{\sigma}_J^4}\left(X_J{}'X_J/n\right) & \mathbf{0}\\ \mathbf{0} & -\frac{2}{\widehat{\sigma}_J^6},\end{pmatrix}'\right) = -\frac{|J|}{\widehat{\sigma}_J^2} - \frac{4}{\widehat{\sigma}_J^2}$$

We conclude that the Kass-Tierney-Kadane expansion of $\mathbb{E}_\pi[\sigma^2|D_n]$ equals

$$\mathbb{E}_\pi[\sigma^2|D_n] = \widehat{\sigma}_J^2 + \frac{1}{n}\left(2\widehat{\sigma}_J^4\left\{\left(\frac{\partial\pi}{\partial\sigma^2}(\widehat{\theta}_J)\right)\cdot\frac{1}{\pi(\widehat{\theta}_J)}\right\} + \widehat{\sigma}_J^2(|J|+4)\right) + O_{\mathbb{P}}\left(\frac{1}{n^2}\right). \qquad \square$$

A.7 Proof of Proposition 3

Proposition 3. *Suppose the true DGP $\mathbb{P}$ satisfies Assumption 1 with parameter $(\beta_0, \sigma_0^2, P_0)$ and associated subset of covariates that are relevant for prediction, J_0. Let Π be a finite collection of priors that satisfy Assumption 2. Further suppose that Π contains at least one prior π^* with $J_0 \subseteq J(\pi^*)$. If*

$$tr\left(n\mathbb{V}_\pi(\beta_{J(\pi)}|D_n)\mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}]\mid D_n\right]\right) = O_{\mathbb{P}}(1), \tag{8}$$

for every prior $\pi \in \Pi$, then

$$\lim_{n\to\infty}\mathbb{P}\left(\exists\pi \in \underset{\pi\in\Pi}{\text{argmin}}\, L^*(\pi, D_n)\ s.t\ J_0 \nsubseteq J(\pi)\right) = 0.$$

Moreover, for any π for which $J_0 \subset J(\pi)$,

$$\lim_{n\to\infty}\mathbb{P}\big(L^*(\pi, D_n) < L^*(\pi_0, D_n)\big) \in (0, 1],$$

where π_0 is any prior for which $J(\pi_0) = J_0$.

Proof. Lemma 1 has shown that the (optimized) posterior mean-squared prediction error for an agent with prior π is

$$L^*(\pi, D_n) = \mathbb{E}_\pi[\sigma^2|D_n] + \text{tr}\left(\mathbb{V}_\pi(\beta_{J(\pi)}|D_n)\mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}]\mid D_n\right]\right).$$

Under the assumptions of Proposition 3, it follows that for any $\pi \in \Pi$:

$$L^*(\pi, D_n) = \mathbb{E}_\pi[\sigma^2|D_n] + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Moreover, Assumptions 1 and 2 imply that the conditions of Lemma 3 are satisfied. Consequently, for any π in the finite collection Π , the term $\mathbb{E}_\pi[\sigma^2|D_n]$ admits the following Kass et al. (1990) expansion:

$$\widehat{\sigma}^2(\pi) + O_{\mathbb{P}}\left(\frac{1}{n}\right), \tag{18}$$

where $\widehat{\sigma}^2(\pi)$ denotes the Maximum Likelihood estimator of σ^2_ϵ according to the linear regression model with covariates $J(\pi)$. Therefore, for any $\pi \in \Pi$ we have

$$L^*(\pi, D_n) = \widehat{\sigma}^2(\pi) + O_{\mathbb{P}}\left(\frac{1}{n}\right). \tag{19}$$

We will use this expansion to prove the two statements of Theorem 3.

Misspecified models never win: We have assumed that the collection Π contains a prior π^* such that $J_0 \subseteq J(\pi^*)$. This prior defeats any other prior π for which $J_0 \nsubseteq J(\pi)$. To see this, note that Equation (19) implies

$$L^*(\pi, D_n) - L^*(\pi^*, D_n) = \widehat{\sigma}^2(\pi) - \widehat{\sigma}^2(\pi^*) + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Under Assumption 1, $\widehat{\sigma}^2(\pi^*) \xrightarrow{p} \sigma^2_0$ (since the larger model includes the relevant covariates). However, since the covariates associated to prior π exclude variables that are relevant for prediction

$$\widehat{\sigma}^2(\pi) - \widehat{\sigma}^2(\pi^*),$$

converges in probability to a strictly positive number (the misspecified model has strictly larger residual variance than the true model). This shows that

$$\lim_{n\to\infty} \mathbb{P}\left(\exists \pi \in \operatorname*{argmin}_{\pi\in\Pi} L^*(\pi, D_n) \text{ s.t } J_0 \nsubseteq J(\pi)\right) = 0.$$

High-dimensional models win with positive probability: For the last part of the theorem, let π_L denote any prior π for which $J_0 \subset J(\pi_L)$ and let π_0 denote any prior for which $J_0 = J(\pi_0)$. From equation (18)

$$L^*(\pi_0, D_n) - L^*(\pi_L, D_n) = \widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L) + O_{\mathbb{P}}\left(\frac{1}{n}\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}(L^*(\pi_L, D_n) < L^*(\pi_0, D_n)) &= \mathbb{P}(n(L^*(\pi_0, D_n) - L^*(\pi_L, D_n)) > 0) \\ &= \mathbb{P}\left(n(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L)) > O_{\mathbb{P}}(1)\right).\end{aligned}$$

where we have used the Kass et al. (1990) expansion in (18). Standard algebra of linear regression—e.g., Equation 5.28 in Greene (2018) and Theorem 5.1 therein—shows that

$$n(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L))/\sigma_0^2 \xrightarrow{d} \zeta,$$

where ζ is a chi-squared random variable with $|J(\pi_L)| - |J_0|$ degrees of freedom.

This shows that

$$\lim_{n\to\infty} \mathbb{P}(L^*(\pi_L, D_n) < L^*(\pi_0, D_n)) \in (0, 1].$$ □

B Supplementary Material

B.1 Proof of Proposition 4

Proposition 4. *Suppose the conditions of Proposition 3 hold. Suppose, in addition, that for any π for which $J_0 \subseteq J(\pi)$,*

$$tr\left(n\mathbb{V}_\pi(\beta_{J(\pi)}|D_n)\mathbb{E}_\pi\left[\mathbb{E}_P[x_{J(\pi)}x'_{J(\pi)}] \mid D_n\right]\right) \xrightarrow{p} \sigma_0^2|J(\pi)|. \tag{20}$$

Then, for any π, π_0 such that $J(\pi_0) = J_0 \subset J(\pi)$,

$$\lim_{n\to\infty} \mathbb{P}\big(L^*(\pi, D_n) < L^*(\pi_0, D_n)\big)$$

converges to the probability that a chi-squared random variable with $|J(\pi)| - |J_0|$ degrees of freedom exceeds

$$2(|J(\pi_0)| - |J_0|) + \left(2\sigma_0^2\left\{\left(\frac{\partial\pi}{\partial\sigma^2}(\beta_0,\sigma_0^2)\right)\cdot\frac{1}{\pi(\beta_0,\sigma_0^2)} - \left(\frac{\partial\pi_0}{\partial\sigma^2}(\beta_0,\sigma_0^2)\right)\cdot\frac{1}{\pi_0(\beta_0,\sigma_0^2)}\right\}\right). \tag{21}$$

Moreover, if the marginal distribution over σ^2 is the same under both π and π_0, the expression in (21) simplifies to

$$2(|J(\pi_0)|-|J_0|)+\left(2\sigma_0^2\left\{\left(\frac{\partial\pi_{\beta|\sigma^2}}{\partial\sigma^2}(\beta_0,\sigma_0^2)\right)\cdot\frac{1}{\pi_{\beta|\sigma^2}(\beta_0,\sigma_0^2)} - \left(\frac{\partial\pi_{0,\beta|\sigma^2}}{\partial\sigma^2}(\beta_0,\sigma_0^2)\right)\cdot\frac{1}{\pi_{0,\beta|\sigma^2}(\beta_0,\sigma_0^2)}\right\}\right).$$

Proof. By Lemma 1 and Equation (20), for any π s.t $J_0 \subseteq \pi$:

$$L^*(\pi, D_n) = \mathbb{E}_\pi[\sigma^2|D_n] + \sigma_0^2(|J(\pi)|)/n + O_{\mathbb{P}}(1)/n.$$

Lemma 3 and Assumption 1 then implies that

$$L^*(\pi_0, D_n) - L^*(\pi, D_n)$$

equals

$$\begin{aligned}
&= \widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi) \\
&+ \frac{1}{n}\left(2\sigma_0^4\left\{\left(\frac{\partial\pi}{\partial\sigma^2}(\beta_0,\sigma_0^2)\right)\cdot\frac{1}{\pi(\beta_0,\sigma_0^2)} - \left(\frac{\partial\pi_0}{\partial\sigma^2}(\beta_0,\sigma_0^2)\right)\cdot\frac{1}{\pi_0(\beta_0,\sigma_0^2)}\right\}\right) \\
&+ 2\sigma_0^2(|J(\pi_0)| - |J|)/n + O_{\mathbb{P}}\left(\frac{1}{n^2}\right).
\end{aligned}$$

As argued in Proposition 3,

$$n(\widehat{\sigma}^2(\pi_0) - \widehat{\sigma}^2(\pi_L))/\sigma_0^2 \xrightarrow{d} \zeta,$$

where ζ is a chi-squared random variable with $|J(\pi_L)| - |J_0|$ degrees of freedom. The result in (21) then follows. To verify the last equation write:

$$\pi(\beta, \sigma^2) = \pi_{\beta|\sigma^2}(\beta, \sigma^2) \cdot \pi_{\sigma^2}(\sigma^2).$$

Using the chain rule and the fact that the marginal distribution of σ^2 is the same under both π and π_0 gives the desired result. □

C Competing Factor Models: Additional details

C.1 Prior hyper-parameters

We have assumed that each agent has a prior of the form

$$\beta|\sigma^2 \sim \mathcal{N}(0, (\sigma^2/\gamma_0 k)\mathbb{I}_k), \quad \sigma^2 \sim \text{Inv-Gamma}(a_0, b_0). \tag{22}$$

In this section we explain how to choose the prior hyper-parameters (a_0, b_0, γ_0). The parameters (a_0, b_0) are common for all agents, but we allow the parameter γ_0 to depend on the agent's relevant covariates.

C.1.1 Choosing γ_0

The prior loss for an agent with prior π and model J is

$$\underbrace{\mathbb{E}_\pi[\sigma^2]}_{\text{Prior Model Fit}} + \underbrace{\mathbb{E}_\pi[\sigma^2]\text{tr}\left(\mathbb{E}_P[x_J x_J']\right)/(\gamma_0 k)}_{\text{Prior Model Uncertainty}}.$$

We take P to equal the empirical distribution of the covariates and set:

$$\gamma_0 = \text{tr}\left(\mathbb{E}_P[x_J x_J']\right)/k. \tag{23}$$

This choice of γ_0 has two justifications. First, it guarantees that all agents have the same prior loss—provided (a_0, b_0) are common among them. Second, it implies that both model fit and model uncertainty contribute equally to the prior loss.

C.1.2 Choosing (a_0, b_0)

Maximizing the marginal likelihood of the data is a common strategy for choosing hyper-parameters in Bayesian Linear Regression; see for example Chapter 3.5 of Bishop (2006). Let X denote the $n \times k$ matrix containing the k covariates for the n observations in the sample. We fix this matrix and analyze the distribution of

$$Y|X, a_0, b_0, \gamma_0, \tag{24}$$

Algebra shows that

$$Y|X, a_0, b_0, \gamma_0, \sigma^2 \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2\left(\mathbb{I}_n + \frac{XX'}{\gamma_0 k}\right)\right). \tag{25}$$

Since $\sigma^2|X, a_0, b_0, \gamma_0$ is Inv-Gamma(a_0, b_0), one can easily compute the joint distribution of $(Y, \sigma^2)|(X, a_0, b_0, \gamma_0)$. Marginalizing σ^2 in such distribution shows that the p.d.f of $Y|X, a_0, b_0, \gamma_0$ equals:

$$\frac{\det\left(\mathbb{I}_n + XX'/\gamma_0 k\right)^{1/2}}{(2\pi)^{n/2}} \frac{b_0^{a_0}}{\Gamma(a_0)} \frac{\Gamma(a_0 + n/2)}{b_n^{a_0+n/2}}, \tag{26}$$

where $\Gamma(\cdot)$ is the Gamma function and

$$b_n = b_0 + \frac{1}{2} Y'\left(\mathbb{I}_n + \frac{XX'}{MSPEk}\right)^{-1} Y.$$

Optimizing (26) with respect to (a_0, b_0) is equivalent to maximizing:

$$a_0 \ln(b_0) + \ln\left(\Gamma(a_0 + n/2)\right) - \ln(\Gamma(a_0)) - (a_0 + n/2)\ln\left(b_n\right). \tag{27}$$

The first order necessary conditions are

$$\begin{aligned} a_0 \quad &: \quad \ln(b_0) + \frac{\Gamma'(a_0 + n/2)}{\Gamma(a_0 + n/2)} - \frac{\Gamma'(a_0)}{\Gamma(a_0)} - \ln(b_n) = 0 \\ b_0 \quad &: \quad \frac{a_0}{b_0} - \frac{(a_0 + n/2)}{b_n} = 0. \end{aligned}$$

A solution to this system of equations must satisfy

$$a_0(b_0) \equiv n \cdot \frac{b_0}{Y'(\mathbb{I}_n + XX'/MSPEk)^{-1} Y}. \tag{28}$$

We plug this equation in (27) and optimize numerically with respect to b_0.

C.2 Further Considerations

C.2.1 Alternative competing models and sample sizes

Figure 3 presented the results of a competition between three different models with two different sample sizes. Figure 4 enriches the baseline comparison in two different dimensions.

The first dimension is to allow for six models. We add the market model (Jensen et al., 1972); the five-factor model recently suggested by Fama and French (2015) (which, relative to the three-factor model adds a 'robust minus weak' profitability factor and a 'conservative minus aggressive' investment factor); and a 42-factor model selected using the recursive double-selection procedure in Feng et al. (2020).

The second dimension is to allow for two additional sample sizes: 175 (5×5) bivariate-sorted portfolios and 1,825 (5×5) bivariate-sorted portfolios that sort based on the subset of the 115 factors that have at least 10 stocks on each quintile cell.[30]

The results of the competition are consistent with what we report in Figure 3, but with some caveats. The low-dimensional models (3 and 5 factors) still perform better than the high-dimensional models (139 and 150 factors) with small samples ($N = 25$ and $N = 175$). However, the Fama and French (2015) five-factor model has a slight edge over the three factor model and market model. This is consistent with the simulations results reported in Figure 3 in the paper. Also, we note that with 175 data points, all the competing models have a similar subjective posterior mean-squared prediction error.

C.2.2 Randomizing selection of portfolios

In our analysis thus far, we have used the 25 bivariate-sorted portfolios on size and book-to-market of Fama and French (1993) as our small sample size. While this reflects the standard choice of test assets in the literature, the superior performance of the three and five factor models relative to high-dimensional models may be specific to the set of test assets considered. Kozak et al. (2018) observes that the Fama and French three factors are similar to the first three principal components of the 25 size and book-to-market portfolios. The three and five factor models may perform well on the small sample size simply because they adequately summarize cross-sectional variation in the 25 size and book-to-market portfolios.

To examine the robustness of our findings to the choice of test asset portfolios, we construct 5,000 simulated datasets, each dataset consisting of randomly chosen portfolios up to a given sample size. For each sample size, we then compute the fraction of times a model

[30]The use of the 175 (5×5) bivariate-sorted portfolios is standard in the literature. They are obtained by doing bivariate sorting using size and each of the following seven variables: book-to-market ratio, Market Beta, "robust minus weak", "conservative minus aggressive", 1-month momentum, 6-month momentum, and 36-month momentum.

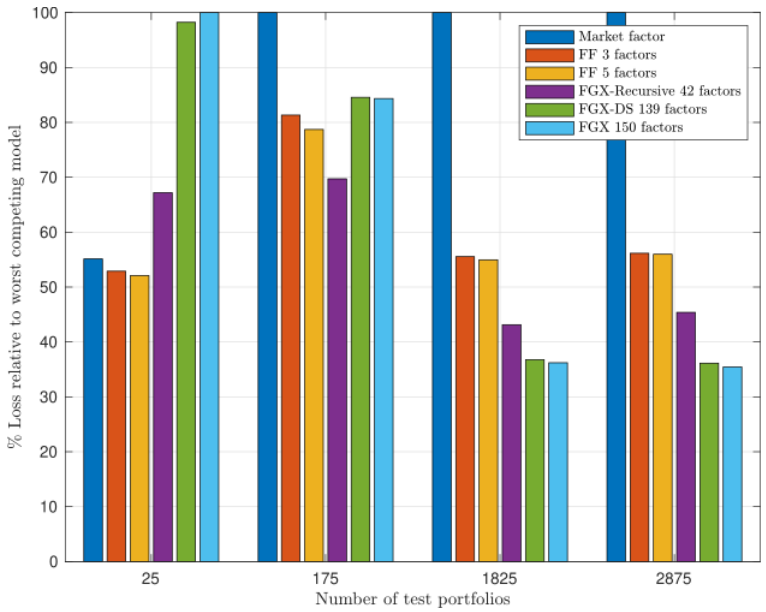

Figure 4: This figure shows the subjective posterior mean squared prediction error of competing models relative to worst model, by number of test portfolios. FF 3 factors refer to the excess market return, Small Minus Big, and High Minus Low factors (Fama and French, 1993), FF 5 factors refer to the FF 3 factors, Conservative Minus Aggressive, and Robust Minus Weak factors Fama and French (2015), FGX-Recursive 42 factors are the factors selected in Feng et al. (2020) (Section C) using a recursive selection procedure, FGX-DS 139 factors are the benchmark factors selected by Feng et al. (2020) using the double-selection method, and FGX 150 factors are all the factors in the factor library of Feng et al. (2020).

achieves the lowest subjective posterior mean-squared forecast error. Figure 5 shows the simulation results. We find that the results remain robust to the choice of test asset portfolios. When $n = 25$, the Fama and French three-factor model prevails 90 percent of the time. As n grows, we see "waves" of larger models performing better, with the 42 factor model out-performing other models for sample sizes between $n = 225$ and 750 and with the largest models (150 factors) prevailing when n is larger than $1,525$.

C.2.3 Competing models over time

The test portfolios used in the exercises above are all constructed by sorting on size and some other factor. Since we have the publication data for each factor, we can easily describe the evolution of the number of available test portfolios over time. We report this in Figure 6, starting from 1976.

We also conduct two additional exercises. First, we consider a sample of only the 25 (5×5) bivariate-sorted portfolios, but consider competition between the three-factor model and the factor zoo available each year. Second, we consider a similar competition, but now

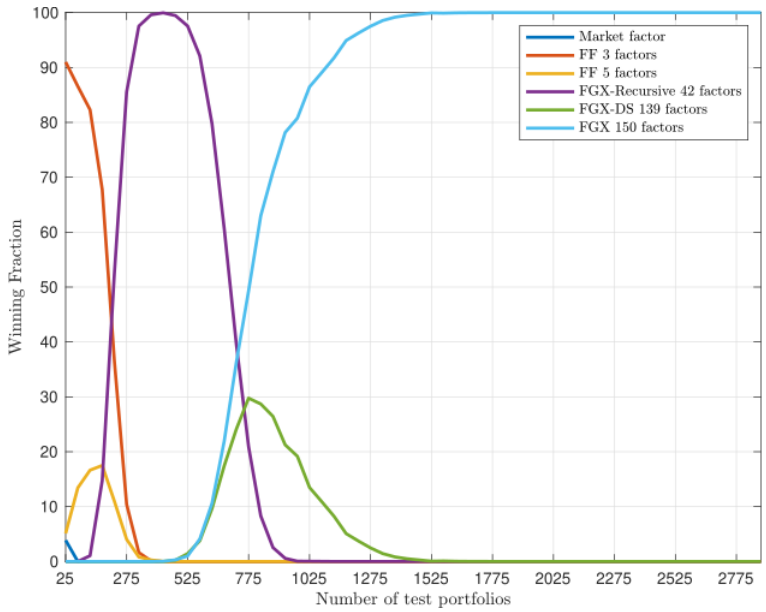

Figure 5: This figure shows winning rates for the competing models over different sample sizes from $n = 25$ to 2875. For each sample size, we construct 5,000 simulated datasets by randomly sampling portfolios without replacement up to the given sample size. The winning rate of a model is computed as the fraction of times the model achieves the lowest subjective posterior mean squared prediction error among all competing models.

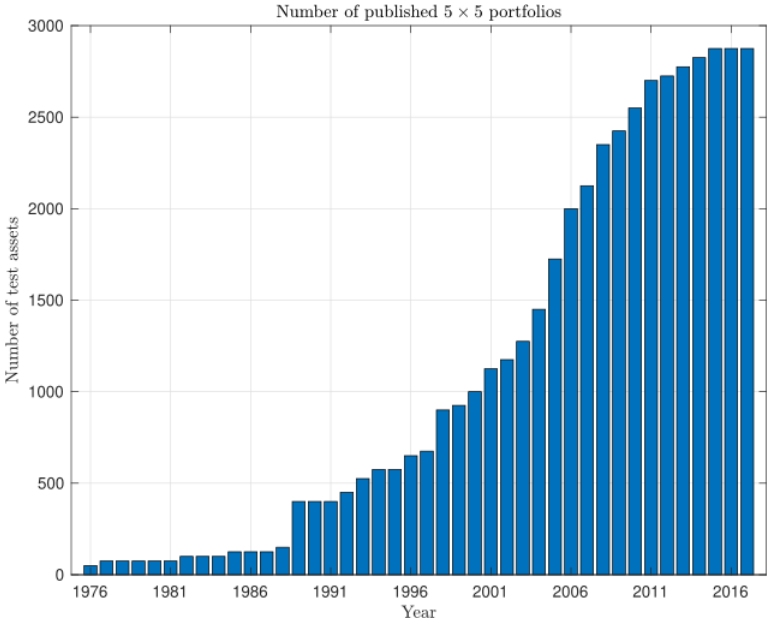

Figure 6: This figure reports the number of 5×5 bivariate-sorted portfolios available as test assets each year for the sample period from 1976 to 2017, based on factors from the Feng et al. (2020) factor library that are published up to the given year.

we assume that the available sample consists of all portfolios published up to a given year. With the small sample, the three-factor model consistently performs better than the factor zoo, with relative performance improving as the size of the factor zoo increases over time. Conversely, when evaluating the models on all available portfolios up to a given time, we see that the factor zoo mostly performs better than the three-factor model. We note that in 1976, both models have similar subjective posterior mean-squared prediction error, albeit with a slight edge for the three-factor model.

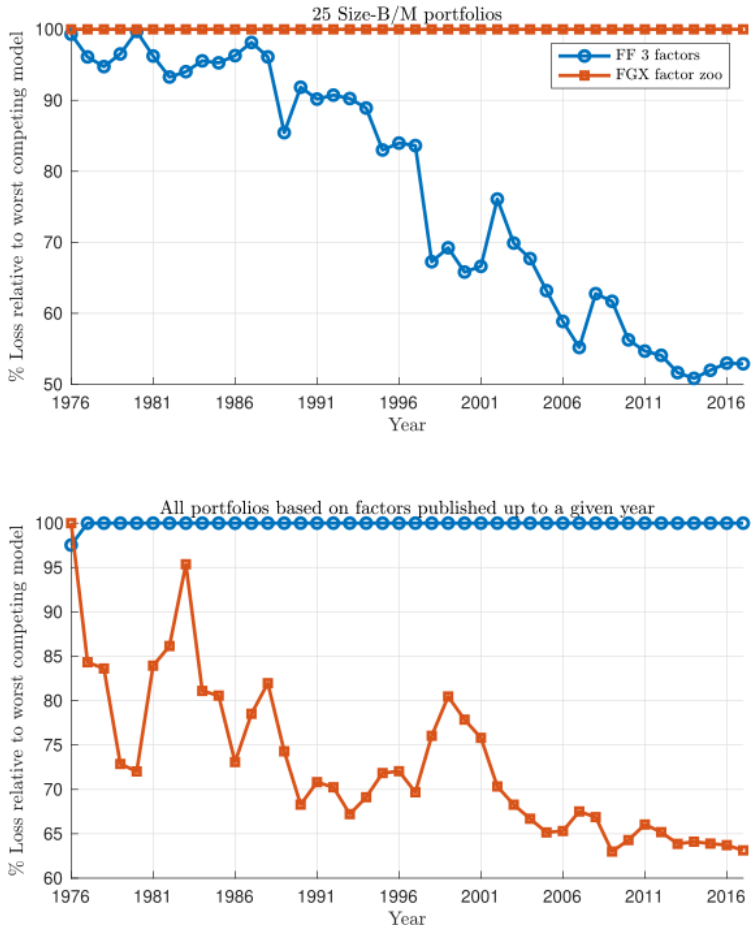

Figure 7: Subject posterior mean squared prediction error of competing models relative to the worst model, over the sample period from 1976 to 2017. For each year, we compute the relative loss using data on returns available up to the given year. The top panel reports the relative loss of the competing models evaluated on 25 size and book-to-market bivariate sorted portfolios, whereas the bottom panel reports the relative loss evaluated on the size and book-to-market bivariate sorted portfolios and 5×5 portfolios based on factors in the Feng et al. (2020) factor library published up given year. The FF3 model includes the market, size, and book-to-market ratio factors. The FGX model includes the FF3 factors as well as all other factors published up to a given year.

C.3 Additional Tables

This table reports the filtering criteria used to select the sample of portfolios. The second column (Number.Factors) states the number of long-short factors that can be constructed from the selected portfolios, and the third column (Number.Portfolios) states the corresponding number of portfolios chosen after the criteria is applied. Selected portfolios from Kenneth French's website includes 25 potfolios sorted by size and book-to-market ratio, 25 portfolios sorted by size and beta, 25 portfolios sorted by size and operating profitability, 25 portfolios sorted by size and investment, 25 portfolios sorted by size and short-term reversal on prior (1-1) return, 25 portfolios sorted by size and momentum on prior (2-12) return, and 25 portfolios sorted by size and long-term reversal on prior (13-60) return.

Filtering criteria	**Number.Factors**	**Number.Portfolios**
Factor library of Feng et al. (2020)	150	
Factors with 5x5 bivariate-sorted portfolios available	115	2875
Factors with 5x5 bivariate-sorted portfolios with at least 10 stocks	73	1825
Selected 5x5 bivariate-sorted portfolios from Kenneth French's website	7	175
5x5 portfolios sorted by size and book-to-market ratio	1	25

This table replicates the list of 150 factors and various descriptive statistics from the factor library in the Appendix of Feng et al. (2020). Of the 150 factors, 15 factors come from publicly available sources from the respective authors' websites, and the remaining 135 factors are constructed from long-short value weighted 3×2 bivariate-sorted portfolios. See Feng et al. (2020) Section II.A.1 for details on factor construction. Asterisks on IDs indicate the 115 factors where 5×5 bivariate sorted portfolios are also available. The values in column $N \geq 10$ equal 1 if there are at least 10 stocks in each of the 5×5 bivariate-sorted portfolios, 0 if there are less than 10 in any of the portfolios, and NA if data is not available. The column Avg.Ret is the monthly average return of the traded factors, Year.Pub is the year of publication, and Authors are the authors of the respective papers.

ID	Description	$N \geq 10$	Avg.Ret	Year.Pub	Authors
1	Excess Market Return	NA	0.64%	1972	Black and Jensen and Scholes
2*	Market Beta	0	−.08%	1973	Fama and Macbeth
3*	Earnings to price	1	0.28%	1977	Basu
4	Dividend to price	NA	0.01%	1979	Litzenberger and Ramaswamy
5*	Unexpected quarterly earnings	1	0.12%	1982	Rendelman and Jones and Latane
6	Share price	NA	0.02%	1982	Miller and Scholes
7*	Long-Term Reversal	1	0.34%	1985	De Bondt and Thaler
8*	Leverage	0	0.21%	1988	Bhandari
9*	Cash flow to debt	0	−.09%	1989	Ou and Penman
10*	Current ratio	0	0.06%	1989	Ou and Penman
11*	% change in current ratio	1	0.00%	1989	Ou and Penman
12*	% change in quick ratio	1	−.04%	1989	Ou and Penman
13*	% change sales-to-inventory	1	0.17%	1989	Ou and Penman
14*	Quick ratio	1	−.02%	1989	Ou and Penman
15*	Sales to cash	1	0.01%	1989	Ou and Penman
16*	Sales to inventory	1	0.09%	1989	Ou and Penman
17*	Sales to receivables	1	0.14%	1989	Ou and Penman
18*	Bid-ask spread	0	−.04%	1989	Amihud and Mendelson
19*	Depreciation / PP&E	1	0.11%	1992	Holthausen and Larcker
20*	% change in depreciation	1	0.08%	1992	Holthausen and Larcker
21	Small Minus Big	NA	0.21%	1993	Fama and French
22*	High Minus Low	1	0.28%	1993	Fama and French

ID	Description	$N \geq 10$	Avg Ret	Year.Pub	Authors
23*	1-month momentum	1	0.15%	1993	Jegadeesh and Titman
24*	6-month momentum	0	0.21%	1993	Jegadeesh and Titman
25*	36-month momentum	0	0.09%	1993	Jegadeesh and Titman
26*	Sales growth	1	0.04%	1994	Lakonishok and Shleifer and Vishny
27*	Cash flow-to-price	1	0.31%	1994	Lakonishok and Shleifer and Vishny
28	New equity issue	NA	0.10%	1995	Loughran and Ritter
29	Dividend initiation	NA	−.03%	1995	Michaely and Thaler and Womack
30	Dividend omission	NA	−.08%	1995	Michaely and Thaler and Womack
31*	Working capital accruals	1	0.22%	1996	Sloan
32*	Sales to price	0	0.35%	1996	Barbee and Mukherji and Raines
33*	Capital turnover	1	−.01%	1996	Haugen and Baker
34*	Momentum	1	0.63%	1997	Carhart
35*	Share turnover	0	−.02%	1998	Datar and Naik and Radcliffe
36*	% change in gross margin - % change in sales	1	−.05%	1998	Abarbanell and Bushee
37*	% change in sales - % change in inventory	1	0.14%	1998	Abarbanell and Bushee
38*	% change in sales - % change in A/R	1	0.14%	1998	Abarbanell and Bushee
39*	% change in sales - % change in SG&A	1	0.09%	1998	Abarbanell and Bushee
40*	Effective Tax Rate	0	−.04%	1998	Abarbanell and Bushee
41*	Labor Force Efficiency	1	−.03%	1998	Abarbanell and Bushee
42	Ohlson's O-score	NA	0.05%	1998	Dichev
43*	Altman's Z-score	0	0.20%	1998	Dichev
44*	Industry adjusted % change in capital expenditures	1	0.10%	1998	Abarbanell and Bushee
45	Number of earnings increases	NA	0.01%	1999	Barth and Elliott and Finn
46*	Industry momentum	1	0.01%	1999	Moskowitz and Grinblatt
47	Financial statements score	NA	0.08%	2000	Piotroski
48*	Industry-adjusted book to market	0	0.22%	2000	Asness and Porter and Stevens
49*	Industry-adjusted cash flow to price ratio	1	0.26%	2000	Asness and Porter and Stevens
50*	Industry-adjusted change in employees	1	−.01%	2000	Asness and Porter and Stevens
51	Industry-adjusted size	NA	0.36%	2000	Asness and Porter and Stevens

ID	Description	$N \geq 10$	Avg	Year.Pub	Authors
52	Dollar trading volume	NA	0.38%	2001	Chordia and Subrahmanyam and Anshuman
53	Volatility of liquidity (dollar trading volume)	NA	0.20%	2001	Chordia and Subrahmanyam and Anshuman
54*	Volatility of liquidity (share turnover)	0	0.02%	2001	Chordia and Subrahmanyam andAnshuman
55*	Advertising Expense-to-market	0	−0.03%	2001	Chan and Lakonishok and Sougiannis
56*	R&D Expense-to-market	0	0.34%	2001	Chan and Lakonishok and Sougiannis
57*	R&D-to-sales	1	0.06%	2001	Chan and Lakonishok and Sougiannis
58*	Kaplan-Zingales Index	0	0.22%	2001	Lamont and Polk and Saa-Requejo
59*	Change in inventory	1	0.18%	2002	Thomas and Zhang
60*	Change in tax expense	1	0.09%	2002	Thomas and Zhang
61	Illiquidity	NA	0.34%	2002	Amihud
62	Liquidity	NA	0.38%	2003	Pastor and Stambaugh
63*	Idiosyncratic return volatility	0	0.07%	2003	Ali and Hwang and Trombley
64*	Growth in long term net operating assets	1	0.22%	2003	Fairfield and Whisenant and Yohn
65*	Order backlog	0	0.05%	2003	Rajgopal and Shevlin and Venkatachalam
66*	Changes in Long-term Net Operating Assets	1	0.24%	2003	Fairfield and Whisenant and Yohn
67*	Cash flow to price ratio	1	0.27%	2004	Desai and Rajgopal and Venkatachalam
68	R&D increase	NA	0.06%	2004	Eberhart and Maxwell and Siddique
69*	Corporate investment	1	0.13%	2004	Titman and Wei and Xie
70*	Earnings volatility	1	0.10%	2004	Francis and LaFond and Olsson and Schipper
71*	Abnormal Corporate Investment	1	0.13%	2004	Titman and Wei and Xie
72*	Net Operating Assets	1	0.31%	2004	Hirshleifer and Hou and Teoh and Zhang
73*	Changes in Net Operating Assets	1	0.14%	2004	Hirshleifer and Hou and Teoh and Zhang
74*	Tax income to book income	1	0.14%	2004	Lev and Nissim
75*	Price delay	0	0.07%	2005	Hou and Moskowitz
76	years since first Compustat coverage	NA	0.01%	2005	Jiang and Lee and Zhang
77*	Growth in common shareholder equity	1	0.15%	2005	Richardson and Sloan and Soliman and Tuna

ID	Description	$N \geq 10$	Avg Ret	Year.Pub	Authors
78*	Growth in long-term debt	1	0.06%	2005	Richardson and Sloan and Soliman and Tuna
79*	Change in Current Operating Assets	1	0.19%	2005	Richardson and Sloan and Soliman and Tuna
80*	Change in Current Operating Liabilities	1	0.03%	2005	Richardson and Sloan and Soliman and Tuna
81	Changes in Net Non-cash Working Capital	NA	0.11%	2005	Richardson and Sloan and Soliman and Tuna
82*	Change in Non-current Operating Assets	0	0.21%	2005	Richardson and Sloan and Soliman and Tuna
83*	Change in Non-current Operating Liabilities	1	0.04%	2005	Richardson and Sloan and Soliman and Tuna
84*	Change in Net Non-current Operating Assets	0	0.23%	2005	Richardson and Sloan and Soliman and Tuna
85*	Change in Net Financial Assets	0	0.23%	2005	Richardson and Sloan and Soliman and Tuna
86	Total accruals	NA	0.19%	2005	Richardson and Sloan and Soliman and Tuna
87	Change in Short- term Investments	NA	−.03%	2005	Richardson and Sloan and Soliman and Tuna
88*	Change in Financial Liabilities	1	0.18%	2005	Richardson and Sloan and Soliman and Tuna
89*	Change in Book Equity	0	0.17%	2005	Richardson and Sloan and Soliman and Tuna
90	Financial statements score	NA	0.17%	2005	Mohanram
91*	Change in 6-month momentum	0	0.21%	2006	Gettleman and Marks
92*	Growth in capital expenditures	1	0.14%	2006	Anderson and Garcia-Feijoo
93*	Return volatility	0	−.02%	2006	Ang and Hodrick and Xing and Zhang
94*	Zero trading days	0	−.05%	2006	Liu

ID	Description	$N \geq 10$	Avg Ret	Year.Pub	Authors
95*	Three-year Investment Growth	1	0.11%	2006	Anderson and Garcia-Feijoo
96*	Composite Equity Issuance	0	−.01%	2006	Daniel and Titman
97*	Net equity finance	0	0.08%	2006	Bradshaw and Richardson and Sloan
98*	Net debt finance	1	0.17%	2006	Bradshaw and Richardson and Sloan
99*	Net external finance	1	0.22%	2006	Bradshaw and Richardson and Sloan
100*	Revenue Surprises	1	0.05%	2006	Jegadeesh and Livnat
101*	Industry Concentration	1	0.03%	2006	Hou and Robinson
102	Whited-Wu Index	NA	−.02%	2006	Whited and Wu
103*	Return on invested capital	0	0.18%	2007	Brown and Rowe
104*	Debt capacity/firm tangibility	1	0.05%	2007	Almeida and Campello
105*	Payout yield	0	0.16%	2007	Boudoukh and Michaely and Richardson and Roberts
106*	Net payout yield	0	0.16%	2007	Boudoukh and Michaely and Richardson and Roberts
107	Net debt-to-price	NA	0.02%	2007	Penman and Richardson and Tuna
108*	Enterprise book-to-price	0	0.14%	2007	Penman and Richardson and Tuna
109	Change in shares outstanding	NA	0.24%	2008	Pontiff and Woodgate
110*	Abnormal earnings announcement volume	1	−.08%	2008	Lerman and Livnat and Mendenhall
111*	Earnings announcement return	1	0.02%	2008	Kishore and Brandt and Santa-Clara and Venkatachalam
112*	Seasonality	0	0.16%	2008	Heston and Sadka
113*	Changes in PPE and Inventory-to-assets	1	0.19%	2008	Lyandres and Sun and Zhang
114*	Investment Growth	1	0.17%	2008	Xing
115*	Composite Debt Issuance	1	0.08%	2008	Lyandres and Sun and Zhang
116	Return on net operating assets	NA	0.09%	2008	Soliman
117*	Profit margin	0	0.02%	2008	Soliman
118	Asset turnover	NA	0.06%	2008	Soliman
119*	Industry-adjusted change in asset turnover	1	0.14%	2008	Soliman
120*	Industry-adjusted change in profit margin	1	−.01%	2008	Soliman
121*	Cash productivity	1	0.27%	2009	Chandrashekar and Rao
122	Sin stocks	NA	0.44%	2009	Hong and Kacperczyk

ID	Description	$N \geq 10$	Avg Ret	Year.Pub	Authors
123*	Revenue surprise	0	0.12%	2009	Kama
124*	Cash flow volatility	0	0.20%	2009	Huang
125*	Absolute accruals	1	−.05%	2010	Bandyopadhyay and Huang and Wirjanto
126*	Capital expenditures and inventory	1	0.19%	2010	Chen and Zhang
127*	Return on assets	0	−.09%	2010	Balakrishnan and Bartov and Faurel
128*	Accrual volatility	0	0.19%	2010	Bandyopadhyay and Huang and Wirjanto
129*	Industry-adjusted Real Estate Ratio	0	0.11%	2010	Tuzel
130*	Percent accruals	1	0.16%	2011	Hafzalla and Lundholm and Van Winkle
131*	Maximum daily return	0	0.00%	2011	Bali and Cakici and Whitelaw
132*	Operating Leverage	1	0.20%	2011	Novy-Marx
133*	Inventory Growth	1	0.13%	2011	Belo and Lin
134*	Percent Operating Accruals	1	0.15%	2011	Hafzalla and Lundholm and Van Winkle
135*	Enterprise multiple	0	0.11%	2011	Loughran and Wellman
136*	Cash holdings	1	0.13%	2012	Palazzo
137	HML Devil	NA	0.23%	2013	Asness and Frazzini
138*	Gross profitability	1	0.15%	2013	Novy-Marx
139*	Organizational Capital	1	0.21%	2013	Eisfeldt and Papanikolaou
140	Betting Against Beta	NA	0.91%	2014	Frazzini and Pedersen
141	Quality Minus Junk	NA	0.43%	2014	Asness and Frazzini and Pedersen
142*	Employee growth rate	1	0.08%	2014	Bazdresch and Belo and Lin
143*	Growth in advertising expense	0	0.07%	2014	Lou
144	Book Asset Liquidity	NA	0.09%	2014	Ortiz-Molina and Phillips
145*	Robust Minus Weak	1	0.34%	2015	Fama and French
146*	Conservative Minus Aggressive	1	0.26%	2015	Fama and French
147	HXZ Investment	NA	0.34%	2015	Hou and Xue and Zhang
148	HXZ Profitability	NA	0.57%	2015	Hou and Xue and Zhang
149	Intermediary Risk Factor	NA	0.15%	2016	He and Kelly and Manela
150	Convertible debt indicator	NA	0.11%	2016	Valta