The Double-edged Sword of Data Mining: Implications on Asset Pricing and Information Efficiency

This draft: November 10, 2023

Abstract

Does data mining always increase price efficiency? Not necessarily. I incorporate data mining into a standard asset pricing model and identify a novel cost of complexity that arises endogenously from data mining. When a data miner explores alternative data, she faces a scarcer training history relative to potential predictors (increasing complexity) and an increasing difficulty in extracting useful signals (decreasing return in data efficacy). The cost of complexity and decreasing return in data efficacy together imply a finite optimal data mining level, such that excess data mining will lead to lower price informativeness. Empirically, I provide evidence of decreasing return in data efficacy in the context of the "factor zoo", and I show that the release of satellite data reduces price informativeness in a difference-in-difference setting.

JEL Code: G11, G12, G14

Keywords: Data mining, Price informativeness, Cost of complexity, Factor zoo, Alternative Data

1 Introduction

In recent years, alternative data sets have become more widely used by investors to make predictions on future asset returns. Asset managers enhance their predictor set by sourcing unconventional data, including news articles, announcement transcripts, and satellite images, in addition to traditional signals from past price patterns and firm fundamentals. The prevailing notion is that data mining efforts enable investors to derive a more precise signal from these predictors, thus enhancing price discovery and improves price efficiency (see Bai et al. (2016); Dávila and Parlatore (2023), among others).

However, in this paper I show that data mining could lead to lower price informativeness. This surprising result arises from the interaction between two realistically motivated model features: *complexity* and *data efficacy*. On the one hand, in the real world investor have only a finite history to estimate the statistical relationship between predictors and future payoff. As the set of predictor expands, insufficient training history introduces complexity into the investor's learning problem. On the other hand, investor find it increasingly challenging to extract truly useful signals from newly acquired data sets, resulting in a decreasing marginal benefit of new predictor.

In this paper, I study a simple model where a representative investor tries to estimate future asset payoffs and incorporates the estimation into prices, with the presence of complexity and decreasing return in data efficacy. Importantly, I assume that the investor takes a "data-driven" approach in estimating payoffs, such that the investor's prior belief puts equal weight on all signals she has. With decreasing return in data efficacy in the signal discovery process, an equal prior weight means the investor pays equal attention to less informative signals as the more informative ones. When training history is sufficient, such equal attention to all predictors doesn't have an effect on the posterior estimates, because there's enough data for the investor to figure out the true predictive relationships. However, when there's insufficient training data, complexity creates limits to learning that prevents precise estimation of the true predictive relationship. In the model, I illustrate that with decreasing return in data efficacy, for the investor with equal prior on all signals, the marginal cost of data mining (higher complexity) eventually dominates the marginal benefit (better approximation of true DGP), leading to a decline in price informativeness.

In the real world, investors must estimate a statistical model to predict future asset payoff from their sets of predictors. With an increasing number of predictors, they grapple with models featuring heavier parameterization. Unlike other machine learning and AI domains, such as auto-driving cars and natural language processing (NLP), where training data is abundant and easy to generate, obtaining additional time series return data for return prediction is not easily achievable. The historical data for training the statistical model thus remains limited, at least in the near future[1]. Additionally, most alternative data sets have a relatively short historical span. As more potential predictors emerge, the training data set becomes increasingly insufficient relative to the size of signals. Thus, data mining elevates the "complexity" of investor's estimation problem.

At the same time, in the real world the newly acquired alternative data is often less useful than the old ones in terms of predictive power[2]. For instance, investors have long explored accounting-based variables like valuation ratios or profitability to predict stock return. These variables are evident predictors because there's direct economic connection between these fundamentals and a firm's future return[3]. In contrast, many alternative data sets, such as web traffic, satellite images, and transcript data, are more challenging to process, have a lower signal-to-noise ratio, and only have an indirect connection to future stock returns. As a result, the "data efficacy" of investor's information set might not scale up linearly with the size of the data set.

In Section 2, I present empirical evidence on the decreasing return in data efficacy within the context of the "factor zoo" and Stochastic Discount Factor (SDF) estimation. Regarding the stock-level return predictors discovered in the previous literature, I demonstrate that the marginal benefit (in terms of SDF Sharpe ratio) declines as more raw predictors are used to construct an out-of-sample SDF estimate. Additionally, employing a neural network to introduce nonlinearity, I show that the out-of-sample Sharpe ratio eventually declines with the inclusion of more predictors.

It remains unclear how the quest of alternative data affects asset prices in the presence of these realistic complications. In this paper, I theoretically characterize the impact of data mining on equilibrium asset prices, accounting for insufficient training history and decreasing return in data efficacy. Without loss of generality, my approach assumes the true asset payoff f_{t+1} is driven by a set of P predictive signals linearly, i.e.

$$f_{t+1} = \sum_{i=1}^{P} \beta_i X_{i,t} + \epsilon_{t+1} \tag{1}$$

On the other hand, the investor (data miner) only has access to a data set of P_1 signals. Among these P_1 signals, P_x of them are the truly useful signal X, while the other signals are redundant, though the investor doesn't know their usefulness ex-ante. The investor needs to

[1] For example, at the end of 2022, the CRSP aggregate stock market return has 1164 monthly observations, while the average number of monthly observations for CRSP individual stock is 133. Such point on insufficient training data is also made in Fernández-Villaverde (2021)

[2] In a Bloomberg article, many practitioners suggested that "investing signals are hard to find" among alternative data sets. For example, one fund manager said "I've looked at probably 700 or 800 data sets over the last 10 years and about 90 to 95% of data sets tend to have basic evident biases to them...They don't really deliver the claims the vendor has made".

[3] See Fama and French (1992, 2015); Zhang (2005); Hou et al. (2015).

estimate the predictive relationship β's of these P_1 signals from past realizations of payoff f and observed signals, but can only observe as far back as T periods. I capture the idea of data mining as an increase in investor's signal set P_1, while recognizing the insufficiency in training history by considering T to be relatively small and fixed compared to P_1. See Figure 4 for an illustration of the data mining process.

To represent the notion of training data insufficiency, I adopt the terminology from Kelly et al. (2021) and define $c = P_1/T$ as *complexity*, which captures the heaviness of model parameterization relative to training data size. Since the training data size T doesn't scale easily in the real world, the increase in P_1 during the data mining process effectively increases c linearly. I show that complexity is closely related to the investor's estimation uncertainty. To capture the varying usefulness in data sets acquired in different phases of the data mining process, I introduce $\kappa = P_x/P$ as the measure of *data efficacy*. A higher κ means the data miner's data set contains many useful signals, while a lower κ implies that most acquired data sets are useless. As the data miner expands the data set size P_1 (and equivalently increases c), more true predictive signals are discovered, leading to an increase in data efficacy κ. However, after exploiting the obvious signals, finding additional true signals becomes more difficult, causing the increase in data efficacy to slow down. Therefore, I assume that κ is an increasing concave function in P_1, capturing the decreasing return to scale in the data mining process.

I explore how a representative, risk averse Bayesian data miner forms posterior belief about future payoff and prices assets in this information environment. I assume that the data miner adopts a prior belief that all signals have the same predictability. This formulation mimics a "quant" investor in the real world, who doesn't have special expertise in analyzing any specific subset of signals, but instead takes a data-driven approach and let the data speak about each signal's predictability. In fact, many widely-used machine learning models such as Ridge or LASSO regression can be thought of as generated by the posterior of a data miner with this equal predictability prior.

I theoretically characterize how price volatility, risk premium, and price informativeness responds to an increase in P_1 due to data mining. As a benchmark, I first characterize pricing results when there's sufficient training history, i.e. when T is large relative to P_1 so that $c \approx 0$. In this scenario, the investor's prior inconsequential, and they can estimate the true β without any estimation uncertainty. As a result, the variation in price aligns with the variation in the payoff from the observed true predictors, while redundant signals receive zero coefficient estimates and exert no impact on price. In this case, despite the decreasing return to scale in data efficacy, price informativeness always increases with data mining.

I then depart from the assumption of sufficient training data and examine the scenario with limited training history. Asset pricing moments differ from the benchmark, with the distortion comes from two sources. First, with complexity $c > 0$, the data miner lacks sufficient training data to recover the true model parameters, resulting in noisy forecasts influenced by sampling errors in the training history. Second, higher complexity induces increased shrinkage in the estimator, which increases conditional bias. This shrinkage stems from both *explicit* and *implicit* sources. Explicit shrinkage occurs when the Bayesian investor, with an informative prior, opts for higher shrinkage to address the challenge in parameter estimation posed by an increase in c. Implicit shrinkage emerges when $c > 1$ due to the fundamental difficulty in constructing conditionally unbiased forecasts[4].

The estimation challenge posed by data mining must be weighed against the benefit it brings, namely an improved information set for capturing the variability in future payoffs, to fully characterize how data mining affects price informativeness. I demonstrate that in the presence of decreasing returns to scale in data efficacy, as observed in the real world with the mining of alternative data sets, the advantage derived from expanding the predictor set diminishes. At some point, it will fall below the marginal cost of complexity. I illustrate that with decreasing returns to scale in data efficacy, there exists an optimal finite level of data mining P_1^* that maximizes both individual investor's utility and the overall price informativeness. Further data mining that increases complexity beyond P_1^* fails to justify the higher estimation difficulty and results in diminished price informativeness and investor utility, as measured by the Sharpe ratio. Intuitively, this surprising result can be understood as a result of investor using the data-driven approach when data has diminishing usefulness. When the data miner adopts a prior belief that all predictors are potentially useful, he will train the model trying to figure out the usefulness of signals that are in fact redundant. Such estimation on redundant signals will interfere with learning the useful signals when there's insufficient training data, thus creating decrease in price informativeness.

I also characterize how price volatility and risk premium evolve with the scale of data mining. For price volatility, I show that it increases at the beginning, as the variation gets captured in the true predictors increases. But price volatility eventually decreases as complexity introduces higher shrinkage. As for the risk premium, I demonstrate that it is linked to data miner's perceived risk, which stems from three sources: uncertainty from parameter estimation, unexplained variation due to missing signals, and truly unlearnable residuals. I delineate how it changes with P_1, illustrating that the risk premium decreases more slowly than in the benchmark case due to nonzero estimation uncertainty.

Finally, while a comprehensive empirical evaluation of the impact of data mining on price efficiency is difficult and beyond the scope of this paper, I offer suggestive evidence that empirically indicates that data mining can lead to lower price informativeness. Specifically, I assess the causal effect of mining alternative data on price informativeness by leveraging the

[4] The implicit shrinkage is the key driver to the "benign overfit" or "double decent" phenomenon in large statistical models. See Hastie et al. (2022), Belkin et al. (2019), Ghosh and Belkin (2022) for more detailed discussions on the implicit shrinkage.

release of satellite data as an expansion to investor's predictor set. Employing a difference-in-difference research design, I show that the release of parking lot traffic data on retailers from satellite image reduces price informativeness for the stocks covered by this data set. I also show that this effect is more pronounced for firms with shorter data availability (higher complexity), suggesting insufficient training history is an important channel in which data mining affects price informativeness.

Literature This paper contributes to the growing literature that studies decision making in a high dimensional setting, such as Aragones et al. (2005), Al-Najjar (2009) and Montiel Olea et al. (2022). My paper is closely related to the recent work in Martin and Nagel (2022), who show that cross-sectional return predictability can arise naturally when investors face parameter uncertainty on asset prices in the high-dimensional learning, and Balasubramanian and Yang (2019) who studies trading game in high dimensional setting with uncertainty about other trader's prior. My Bayesian agent setup is similar to Martin and Nagel (2022), but instead of focusing on return predictability from risk neutral agents, I study the general equilibrium pricing when the agent is risk averse, and I study how data mining affects risk premium, price volatility, and price informativeness in the equilibrium. My paper also connects to the literature that emphasizes the role of parameter estimation in shaping asset pricing moments (Lewellen and Shanken (2002), Pastor and Veronesi (2009), Collin-Dufresne et al. (2016)). Many of these models assume low-dimensional data generating process but with regime shifts, such that agents effectively has limited training data in each regime to estimate parameters precisely. On the contrary, my model directly formulates parameter uncertainty when the number of potential signals increases.

This paper also contributes to the study of big data and its impact on asset prices (Goldstein et al. (2021); Farboodi et al. (2022a)). My model provides realistic characterization on the process of data mining, and I explicitly derive the limiting asset pricing moments in closed form. My paper also speaks to information and learning models in finance (e.g. Vives (2010), Farboodi et al. (2022a); Veldkamp (2023b)). In these models, the cost of information acquisition plays an important role. Many studies justifies the cost of information by the cost of purchasing data set, or the burden on information processing. In my paper, the cost of information arises endogenously as data mining process expands, because it makes training data scarcer relative to the number of parameters, which leads to higher estimation difficulty. My paper also connects to the recent literature that tries to value data as an asset (Farboodi et al. (2022b), Veldkamp (2023a)), such that my model suggests that the length of training history is an important determinant of investor's valuation for data. My work also connects to the recent literature that studies the effect of reducing information acquisition cost on price informativeness, such as Banerjee et al. (2018); Dugast and Foucault (2018, 2023). In many works in this literature, reduction in information acquisition cost doesn't necessarily lead to higher price informativeness due to investor's strategic choice when there's heterogeneity in trading motives or properties of signals. This paper identifies a more direct mechanism why data mining might not lead to more efficient price even in the absence of any heterogeneity.

This paper also connects to the emerging literature on applying machine learning models to return predictions (e.g. Gu et al. (2020), Freyberger et al. (2020), Nagel (2021), Kelly and Xiu (2023)). In this literature, researchers propose to use sophisticated nonlinear models to approximate the unknown functional form of asset returns. The virtue of complexity literature (e.g. Kelly et al. (2021), Didisheim et al. (2023)) shows that when expanding model size, the gain from better approximation dominates the cost of estimation uncertainty when the model admits universal approximation property. While I use a similar model setup, my paper studies expanding raw predictor data sets rather than model approximation capacity. The decreasing return to scale in data efficacy contrasts the constant return to scale in modeling capacity as in this literature, and is a key distinction to understand the impact of data mining in the real world.

The paper proceeds as follows. In Section 2, I provide empirical evidence on decreasing return in data efficacy in the context of SDF estimation. In Section 3, I set up the model, introduce the process of data mining, and derive the benchmark asset pricing results when there's sufficient history. In Section 4, I derive the main asset pricing results with insufficient history and decreasing return in data efficacy. In Section 5 I estimate the causal impact of satellite data release on price informativeness. Section 6 concludes. Additional results and all proofs are in appendix.

2 Evidence of Decreasing Return in Data Efficacy

In this section, I offer empirical evidence of decreasing return in data efficacy in the context of a standard empirical problem in asset pricing: estimating the Stochastic Discount Factor (SDF) from US stock returns. I demonstrate that the marginal benefit of data mining (discovering additional predictors), measured by increase in out-of-sample Sharpe ratio, declines as the predictor set expands, indicating a decreasing return in data efficacy.

2.1 Data

I obtain monthly stock characteristics data constructed in Jensen et al. (2022) (JKP henceforth), which includes 153 characteristics for each stock from 1963 to 2019[5]. Since some

[5] The JKP data set is publicly available at https://jkpfactors.com/. It includes NYSE/AMEX/NASDAQ securities with CRSP share code 10, 11 or 12. For our analysis I exclude "nano" stocks, which are stocks with market capitalization less than 1% across all NYSE stocks.

of the JKP characteristics have low coverage, especially in the early parts of the sample, I reduce the 153 characteristics to a smaller set of 130 characteristics with the fewest missing values. I drop stock-month observations for which more than 30% of the 130 characteristics are missing. Next, I cross-sectionally rank-standardize each characteristic and map it to the [-0.5, 0.5] interval, following Gu et al. (2020). The final 130 stock-level characteristics forms the potential predictor set[6].

2.2 Empirical result

I replicate the real-world data mining process by progressively using additional predictors to form factor portfolios and using them to construct the SDF. Specifically, at each time t, I construct factor returns $F_{t+1} = S_t' R_{t+1}$, where S_t is the matrix of cross-sectional values of characteristics at time t. I then build SDF from those factor returns by estimating a Maximum Sharpe Ratio Regression[7]:

$$\begin{aligned} \hat{\lambda}(z) &= \left(zI + \hat{E}[F_t F_t']\right)^{-1} \hat{E}[F_t] \\ &= \arg\min_{\lambda} \left(\sum_{t=1}^{T} (1 - \lambda' F_t)^2 + z||\lambda||^2 \right) \end{aligned} \tag{2}$$

Here, $\hat{\lambda}(z)$ is the SDF weight on factors, T is the estimation period, and $\hat{E}[F_t] = \frac{1}{T}\sum_t F_t$ and $\hat{E}[F_t F_t'] = \frac{1}{T}\sum_t F_t F_t'$ are the sample mean and second moment of the factors. The out-of-sample SDF portfolio is then constructed as

$$\hat{R}_{T+1}^{SDF} = \hat{\lambda}(z)' F_{T+1} \tag{3}$$

Intuitively, this regression finds the combination of factors F_t that behave as closely as possible to a positive constant (in the l_2 sense), which is tantamount to finding the portfolio with the highest Sharpe ratio.

In the first exercise, I gradually increase the size of S_t from 2 to 130 using the raw characteristics from JKP, and estimate the out-of-sample SDF considering a grid of ridge penalties $z \in \{10^n | n \in \{-9, \cdots 3\}\}$. Starting from January 1973, in each month t, I use a rolling window of 120 months to estimate $\hat{\lambda}$ in (2). I then compute the out-of-sample SDF portfolio return in the subsequent month, and from the sequence of out-of-sample monthly SDF returns I compute its Sharpe ratio.

[6] One concern with this data set is that these predictors are selected from previous studies that are shown to strong predictability of future returns, therefore doesn't generalize to the new signals yet to be discovered. In Appendix D.1 I extend the predictor set to include hypothetical signals, and I show that out-of-sample Sharpe ratio will not improve when more signals are included if the training period is short, i.e. there's nontrivial complexity in the estimation problem.

[7] see Didisheim et al. (2023); Kelly and Xiu (2023)

Figure 1 shows the out-of-sample SDF Sharpe ratio as I progressively increase predictors for factor construction. I focus on the best Sharpe ratio across all shrinkage z. To mitigate randomness in the order of addition, I perform 20 random orderings and report their average. The central conclusion of my analysis is that the marginal benefit of using additional predictor diminishes as the predictor size increases: the increase in Sharpe ratio slows once more predictors are discovered, eventually plateauing. In the terminology used throughout the paper, data efficacy exhibits decreasing returns. In Figure 12 and 13 in Appendix D I replicate the exercise for different rolling windows and in different size groups, and I find the decreasing return pattern is robust.

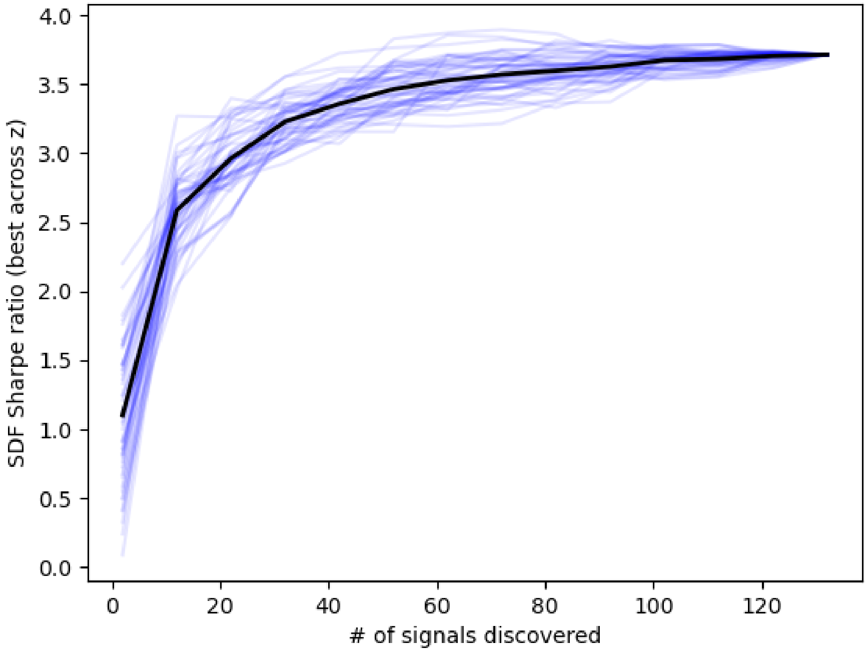


Figure 1: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed by different sizes of predictor sets. I gradually increase the set of JKP predictors used to construct factors and estimate SDF using (2) and (3). I report the highest Sharpe ratio across shrinkage z. Each light blue line represent a random order of discovering predictors, and the black line is the average across random orderings.

In Figure 1, although the gain in Sharpe ratio is decreasing, the general pattern is increasing in the number of raw predictors discovered. This is because the raw JKP predictors are strong predictors of the cross-sectional expected return by previous research. Could it be the case that adding more predictors decreases the SDF performance? I explore this question by adding more "data-mined" variables constructed in Chen et al. (2022) and see how SDF Sharpe ratio changes. These variables are constructed by taking ratios and scaled first differences of Compustat accounting variables and CRSP market equity. I gather 991 value-weighted long short factors built on data-mined signals that have no missing observations and have positive Sharpe ratio in the full sample. This process ensures that the data-mined signals indeed carry predictability of expected return in the cross-section. I then gradually add the data-mined factors on top of the 130 JKP factors to and estimate the SDF Sharpe ratio, following a decreasing order of individual factor's ex-post Sharpe ratio. In other words, I assume that investor discovers the strongest signal first, followed by the second strongest and so forth.

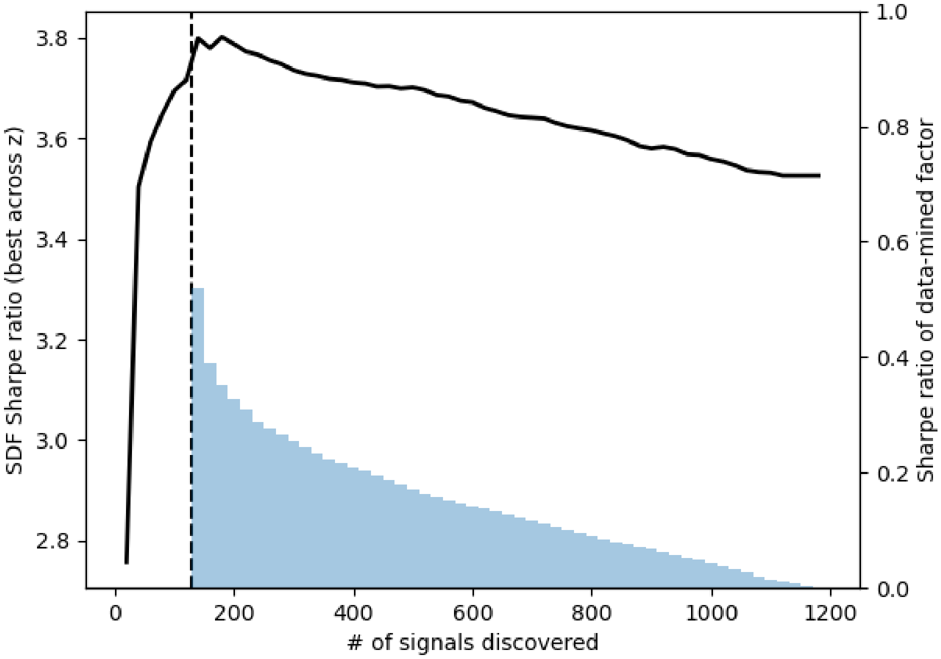


Figure 2: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed from both discovered and hypothetically data-mined factors. The first 130 factors are discovered by previous literature and constructed in JKP. The following factors are data-mined from Chen et al. (2022). The black line (left axis) shows the out-of-sample SDF Sharpe ratio, and the blue bars shows the full-sample Sharpe ratio of each added data-mined factor.

Figure 2 shows how SDF out-of-sample Sharpe ratio changes when we further data-mine accounting variables and include hypothetical factors based on their full-sample Sharpe ratio. As individual data-mined signal each have positive Sharpe ratio, they all contain predictability about future cross-sectional stock returns to some degree. However, when we combine their predictability together and estimate an SDF jointly, the performance starts to decrease when we have more factors in the model. Put it differently, more signals hurts model performance when combined together even when each one of them carries incremental information.

Finally, in the third exercise, rather than directly using a subset of the 130 predictors, I use their nonlinearly transformed counterparts to create factors. This approach can be understood as building a machine learning SDF, where the nonlinear transformations can be viewed as nonlinear basis functions for nonparametric approximators of SDF weights. Following Didisheim et al. (2023) I use random Fourier features (RFF) to transform the raw signals. Specifically, at each time t, I build

$$\tilde{S}_t = [\sin(\gamma S_t \omega_1), \cdots \sin(\gamma S_t \omega_{P/2}), \cos(\gamma S_t \omega_1), \cdots \cos(\gamma S_t \omega_{P/2})] \tag{4}$$

where ω_i is a random vector with same size as S_t drawn from iid normal distributions. Each column in $\tilde{S}_t$ can be thought of as a random linear combination of the raw characteristics S_t fed through the trigonometric activation functions. As before, I progressively increase the set of predictors in S_t from 2 to 130 and for $P \in \{100, 500, 1000\}$ I perform the RFF transformation to generate $\tilde{S}_t$. I then use $\tilde{S}_t$ to generate factors and estimate SDF as described before[8].

Figure 3 shows the out-of-sample Sharpe ratio of machine learning SDF. I arrive at the same conclusion as before: that marginal benefit of additional predictor is decreasing. In fact, for fixed model capacity (fixed P), the marginal benefit can be negative–using more predictors as input to nonlinear transformations eventually harms the out-of-sample Sharpe ratio. This evidence suggest that additional predictors become "redundant" once many predictors are already discovered for models with fixed capacity, and assigning nonzero weights to them impairs model performance. In Figure 14 and 15 in Appendix D I again find the pattern is robust in different rolling windows and size groups.

In conclusion, these empirical results offer evidence of decreasing returns in data efficacy for stock-level predictors: as investors mine additional predictors, the marginal benefit of new predictor decreases, and investors will fail to utilize the incremental information provided by newly-mined predictors, resulting in a decrease in model performance. Motivated by these empirical facts, I then proceed to build a model to study the asset pricing implication of data mining, and in Section 4 I incorporate decreasing return in data efficacy in the model.

3 An Asset Pricing Model with Data Mining

In this section, I introduce an asset pricing model while formulating the idea of data mining. I begin with an economy featuring a single asset and a representative risk-averse Bayesian investor (data miner). The investor needs to forecast asset payoff generated by a high-dimensional set of predictive signals using past realizations of signals and payoffs. The

[8] The parameter γ controls the Gaussian kernel bandwidth in generating random Fourier features. Following Didisheim et al. (2023), I randomly choose γ from the grid $[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$ for each ω_i that I generate. This embeds varying degrees of nonlinearity in the generated features set $\tilde{S}_t$. For each nonlinear feature, I again cross-sectionally rank-standardize it to be in $[-0.5, 0.5]$ interval.

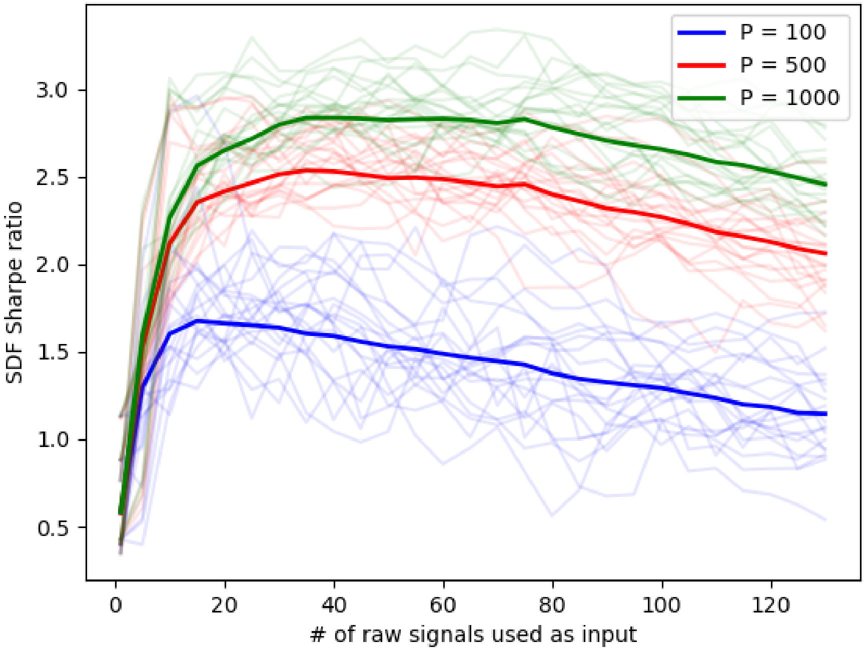


Figure 3: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed by random Fourier features (RFF) of different sizes of predictor sets. I gradually increase the set of JKP predictors used in RFF as in (4) and estimate SDF using (2) and (3) on RFFs with different size P. I average across 20 draws of random weights and report the highest Sharpe ratio across shrinkage z. Each light blue line represent a random order of discovering predictors, and the black line is the average across random orderings.

process of data mining involves expanding the set of signals, which includes both true predictors and redundant predictors, used by the investor to forecast the payoff. I derive the posterior distribution of the asset payoff as a function of the number of predictors and number of training data points. As a benchmark, I characterize the equilibrium risk premium, price volatility, and price informativeness when the investor can observe sufficient history to estimate parameters.

3.1 Environment

Representative investor Time is discrete and there is a single consumption good (dollar) at each date, which serves as numeraire. There are two assets traded in the economy: one risk-free asset with perfectly elastic supply and a gross return which is normalized to 1. There's also a short-lived risky asset with a price p_t at each t, and obtaining one unit of the risky asset at t gives a random payoff f_{t+1} at $t+1$. The supply of the risky asset is fixed at Q. The payoff process is to be specified later and the price is to be determined in the equilibrium.

The risky asset is priced by a short-lived representative investor in the economy, who enters a trade at t and makes consumption at $t+1$. At time t he optimally chooses a position in the risky asset, denoted by q_t, and the remaining wealth in the risk free asset to solve

$$\max_{q_t} E_t^I[U(W_{t+1})] \tag{5}$$

subject to the budget constraint

$$W_{t+1} = W_t + (f_{t+1} - p_t)q_t \tag{6}$$

Here $E_t^I[\cdot]$ denotes the investor's expectation and could be different from the true DGP of the payoff. The equilibrium is a solution q_t such that it maximizes 5 and it clears the risky asset market, i.e. $q_t = Q$. Applying second order approximation to the utility function $U(\cdot)$, one can show that q_t approximately solves a standard mean-variance problem

$$q_t \approx \frac{1}{\rho}\frac{E_t^I[f_{t+1}] - p_t}{V^I[f_{t+1}]} \tag{7}$$

where $\rho = -\frac{W_t U''(W_t)}{U'(W_t)}$ is the relative risk aversion, and $V^I[f_{t+1}]$ is the investor's conditional variance of the payoff. By imposing the market clearing condition, the equilibrium price is determined as

$$p_t = E_t^I[f_{t+1}] - \rho Q V_t^I[f_{t+1}] \tag{8}$$

Assumptions on the Payoff I assume that the random payoff is driven by a potentially high dimensional set of predictors, and the investor needs to estimate the predictive relationship. Specifically, I assume that

$$f_{t+1} = \beta' \tilde{X}_t + \epsilon_{t+1} \tag{9}$$

where $\tilde{X}_t$ is a $P-$dimensional vector of predictors, β is a $P-$ dimensional vector of coefficients that describes the predictive relationship. I denote $\theta_t \equiv \beta' \tilde{X}_t$ as the learnable part of the payoff, and ϵ_{t+1} is the unlearnable part. I make the following assumption governing the true predictive coefficient β, the predictor $\tilde{X}_t$, and unlearnable part ϵ_{t+1}:

Assumption 1. I assume that

1. $\tilde{X}_t$ is a random vector with independent entries that satisfies $E[\tilde{X}_{i,t}] = E[\tilde{X}^3_{i,t}] = 0$ and $E[\tilde{X}^2_{i,t}] = \sigma^2_x$ and finite fourth moment for all $i = 1 \cdots P$.

2. β is random with $i.i.d$ coordinates β_i that are independent of $\tilde{X}_t$ and ϵ_{t+1} such that $E[\beta_i] = 0$ and $E[\beta^2_i] = b_*/P$

3. ϵ_{t+1} is $i.i.d$ with finite fourth moment. $E[\epsilon_{t+1}] = E[\epsilon^3_{t+1}] = 0$ and $E[\epsilon^2_{t+1}] = \sigma^2$

Assumption 1 guarantees that the investor's belief and equilibrium price is well-behaved with large T and P. The first assumption states that the predictors are isotropic, which allows us to derive closed form solution to the equilibrium price. In Appendix B I extend the result to correlated features. The second assumption introduces the randomness of β, which is a device that allows us to solve for price for generic predictive coefficients. Intuitively, one can interpret this condition as the Nature draws true β from a multivariate normal distribution with zero mean and covariance matrix of $\sigma^2_\beta \mathbf{I}_P/P$. The assumption that β has zero mean is inconsequential; one could allow for non-zero mean and restate our analysis in variances instead of second moments. I assume an identity matrix of β to say that the predictability is independent across different predictors. The scalar b_* denotes the L_2 norm of the β vector, i.e. $||\beta||_2 = b_*$. It controls the total amount of variation in payoff that can be learned. Because of the independence of β and $\tilde{X}$, each component $\beta' \tilde{X}_t$ has mean 0 and variance $\frac{b_* \sigma^2_x}{P}$. Thus, the learnable component θ can be viewed as an average of P independent random variables with mean 0 and variance $b_* \sigma^2_x$. By the central limit theorem, as $P \to \infty$, the unconditional (prior) distribution of θ converges to a normal distribution $N(0, b_* \sigma^2_x)$.

3.2 Data Mining and Investor's Learning Problem

The representative investor doesn't observe θ_t. Instead, the investor observes a data set, which consists of a subset of the true predictors and some redundant predictors for past T

periods[9]. Specifically, I assume that at t the investor observes X_τ for $\tau = t - T, t - T + 1, \cdots t - 1$, where $X_\tau = [X_{1,\tau}, W_\tau]$. Here $X_{1,\tau}$ denotes the set of predictors that are included in the true predictor set $\tilde{X}_\tau$. I denote the size of $X_{1,\tau}$ using P_x and obviously $0 \leq P_x \leq P$. In addition, there are predictors W_τ in investor's data set that are not in the true DGP, i.e. they are redundant. I make the following assumption on the distribution of W_t:

Assumption 2. We assume that W_t is a random vector with independent entries that $E[W_{i,t}] = E[W_{i,t}^3] = 0$ and they have the same variance as true predictors for simplicity, i.e. $E[W_{i,t}^2] = \sigma_x^2$.

Assumption 2 states that the redundant signals and true signals have same distributions, and their only difference is whether they are useful in payoff prediction. I denote the size of W_τ using P_w, and I define $P_1 \equiv P_x + P_w$ be the total number of predictors the investor observes. Figure 4 illustrates the sizes of predictors in true DGP and investor's data set.

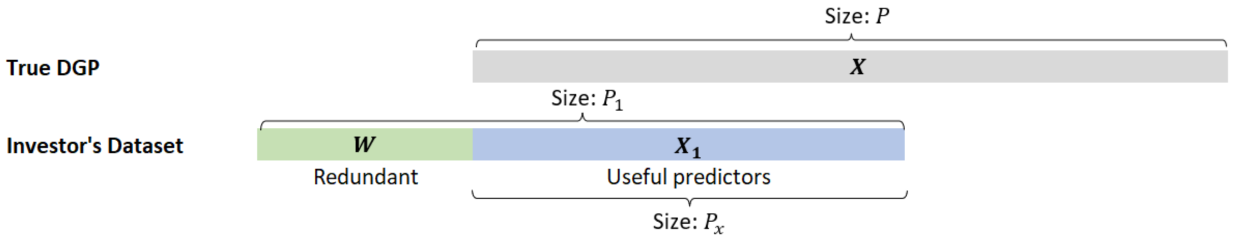


Figure 4: Illustration of payoff process and investor's information set. The gray bar represents a P-dimensional vector $\tilde{X}$ in the true DGP. Investor observes the first P_x predictors of X, which is shown in the blue bar, plus some redundant predictors W with size P_w as in the green bar. The total size of blue and green bar represents the total number of predictors in investor's information set, $P_1 = P_x + P_w$.

This formulation of investor's information set captures the notion of data mining. $\{\tilde{X}_\tau\}_{\tau=t-T}^{t-1}$ can be thought of as the data set provided by a data vendor, which contains possibly relevant predictors for their historical values for T periods. Data mining refers to the process of acquiring new predictors in investor's information set, i.e. an increase in P_1. However, in the real world the investor doesn't know the true data generating process, and thus doesn't know if the newly acquired predictor is a true predictor or a redundant one. In other words, for any X_j in X, the investor doesn't know if it is a true predictor or a redundant one. She thus needs to estimate a predictive relationship $f_{\tau+1} = \beta X_\tau + \epsilon_{\tau+1}$ using $\{f_{\tau+1}, X_\tau\}_{\tau=t-T}^{t-1}$ and make a forecast of $E_t[f_{t+1}]$ and $V_t[f_1]$ using X_t and estimated $\hat{\beta}$.

[9] In this paper I don't study where does investor obtain the data set, and I assume there's no additional cost associated with these data. In practice data sets are costly and are usually supplied by data vendors, who might have a difference incentive than investors. I leave these extensions for future research.

3.2.1 Data efficacy

I then define the concept of data efficacy, which is a key concept in our model and a driver on price efficiency. The data efficacy κ is simply defined as the fraction of the true predictors included in investor's data set.

Definition 3.1. The data efficacy of investor's data set, κ, is defined as

$$\kappa \equiv \frac{P_x}{P} \tag{10}$$

As an investor undertakes data mining which increases P_1, her data set should include more true predictors. Thus one should expect κ to be increasing along with data mining. κ can be further decomposed into two parts:

$$\kappa \equiv \frac{P_x}{P} = \frac{P_x}{P_1}\frac{P_1}{P}$$

Here it's obvious that data efficacy can be driven by two factors. First, it's driven by how useful the data set is, captured by $\frac{P_x}{P_1}$, which measures the proportion of the investor's information set that is truly useful for prediction. A lower $\frac{P_x}{P_1}$ means the investor's information set contains many redundant predictors that are not useful for predicting payoff. Second, it's driven by how well the model capacity compares to the true DGP, captured by $\frac{P_1}{P}$, namely the ratio between the total number of observable predictors and total true parameters. It's easy to see that P_1/P scales linearly with P_1 once the true DGP is fixed. Note that a correct specification means $\kappa = 1$.

As demonstrated in Section 2, discovering true signals becomes more challenging as the observed data set increases in size. Conceptually, as the predictability of obvious predictors saturates, investors need to expand their search for the non-obvious ones. During this search, investors may pickup more redundant signals before encountering a real one. For example, in stock return prediction, starting with price-based and accounting-based signals is obvious, and many of them are robust return predictors. However, once investors move beyond these obvious predictors and search for other predictors in alternative data sets like earnings transcript or social media, the potential signals that can be extracted become enormous. Therefore, it's reasonable to assume that investors need to pass through more redundant signals before finding a new one. For this reason, one should expect the growth rate in κ to decline, i.e. it exhibits decreasing returns. Assumption 3 states the assumption on the evolution of data efficacy κ as a function of data mining scale P_1.

Assumption 3 (Decreasing return in data efficacy). I assume $\kappa'(P_1) > 0$ and $\kappa''(P_1) < 0$

Remark 1 (Decreasing importance in β). In Assumption 1, I assumed predictive coefficient β_i is rotationally symmetric, i.e. all predictors have the same predictability ex-ante. In this case, data efficacy κ can be interpreted as number of true predictors covered in investor's signal set. Section A.1 in Appendix provides another formulation of decreasing returns in κ based on decreasing importance in β. In that formulation, a decreasing return in data efficacy simply means the newly discovered signals carry less predictability than old signals. Both formulations give the same pricing implications.

Remark 2 (Low-dimensional factor structure). The assumption that predictors X_i are distributed i.i.d means there's no factor structure in the true DGP, and dimensionality reduction techniques such as PCA will not solve the high dimensional learning problem in this model. My result can be interpreted as investors trying to predict the idiosyncratic component of the payoff, after common factors are controlled for. It maps naturally to the real world, where sophisticated investors such as hedge funds care about payoffs beyond exposures to common factors. In fact, if the payoff is generated by a low dimensional factor structure and all potential signals are true factors plus noise, then discovering new signals will always lead to higher prediction accuracy and price informativeness because accessing a larger cross-section of signals will allow investors to extract the latent factors more precisely, for example through principal component analysis.

3.2.2 Investor's learning

I assume that the investor is Bayesian and forms a prior regarding the prediction coefficient $\tilde{\beta}$, and makes forecast using the posterior distribution. To insure tractability, I make the following assumption regarding the investor's prior about β:

Assumption 4. We assume that without seeing any data, investor has a prior distribution of $\tilde{\beta}$ that is

$$\beta \sim N(0, \sigma_\beta^2 \mathbf{I}_{P_1}/P_1) \tag{11}$$

Assumption 4 implies that without seeing data, the investor thinks all predictors are equally useful for prediction. This is a natural assumption for a data miner who has know prior knowledge on the usefulness of predictors and just "let the data speak"[10]. σ_β^2 controls the total amount of predictability the investor believes her predictors has. When $\sigma_\beta^2 = 0$, investor thinks all β's are zero and thus the payoff is unpredictable using her data set, whereas when $\sigma_\beta^2 \to \infty$ the investor thinks her predictors are very predictive.

Given the normal prior and the linearity assumption of the payoff, the following proposition characterizes the investor's posterior belief about the payoff

[10] An alternative research doctrine is to use theory to guide empirical model in order to prevent data mining bias, as advocated in Harvey et al. (2016); Harvey (2017). However, recent work by Chen et al. (2022) finds that predictors which peer-reviewed theory argued to be robust still suffers deterioration in predictability out-of-sample.

Proposition 3.1
Given the prior $\beta \sim N(0, \frac{\sigma_\beta^2}{P_1}\mathbf{I}_{P_1})$*, after observing historical data* $X_T = \{X_\tau, f_{\tau+1}\}_{\tau=t-T}^{t-1}$ *and current realization* X_t*, the informed investor holds posterior distribution of* θ_1 *according to*

$$\theta|X_T, f_T, X_t \sim N(\underbrace{X_t'\hat{\beta}}_{E_t^I[\theta_t|\mathcal{I}_I]}, \underbrace{X_t'\hat{V}_\beta X_t}_{V_t^I[\theta_t|\mathcal{I}_I]}) \tag{12}$$

where

$$\begin{aligned}\hat{\beta} &= \left(\frac{P_1\hat{\sigma}^2}{T\sigma_\beta^2}I + \frac{1}{T}X_T'\tilde{X}_T\right)^{-1}\frac{1}{T}X_T'f_T \\ \hat{V}_\beta &= \hat{\sigma}^2\left(\frac{P_1\hat{\sigma}^2}{T\sigma_\beta^2}I + \frac{1}{T}X_T'X_T\right)^{-1}\frac{1}{T^2}X_T'X_T\left(\frac{P_1\hat{\sigma}^2}{T\sigma_\beta^2}I + \frac{1}{T}X_T'X_T\right)^{-1}\end{aligned} \tag{13}$$

and $\hat{\sigma}^2 = \sigma^2 + \frac{1}{T}\sum_\tau \beta_2' X_{2,\tau}$ *where* $X_{2,\tau}$ *and* β_2 *are the true predictors and coefficients in* $\tilde{X}$ *but not observed in* X*. Using the definition of data efficacy in* (10)*, we have*

$$\hat{\sigma}^2 = \sigma^2 + (1-\kappa)b_*\sigma_x^2 \tag{14}$$

Proposition 3.1 shows that the posterior mean of the payoff, which corresponds to the investor's conditional expectation, can be interpreted as a prediction from a Tikhonov-regularized (i.e. ridge) regression with shrinkage parameter $\tau \equiv \frac{P_1\hat{\sigma}^2}{T\sigma_\beta^2}$, for example see Shalev-Shwartz and Ben-David (2014). Indeed, the estimated $\hat{\beta}$ is stabilized by adding a penalty term to the sample covariance matrix, and the prior variance σ_β^2 controls the degree of shrinkage. When $\sigma_\beta \to \infty$, $\hat{\beta}$ becomes the usual OLS estimator. Given the conditional expectation and conditional variance, one can plug them in Eqn. 8 to solve for equilibrium price and other moments. It becomes clear in this formulation that data mining has two effects. The first effect is on the "perceived" unexplanable variation. An increase in P_1 presumably finds more true predictors, and thus reducing the perceived unexplanable variance $\hat{\sigma}^2$. The second effect is on the estimation: as P_1/T increases, the $\hat{\beta}$ estimates deviates more from the OLS estimates and it introduces more bias. I will characterize how these two effects matter for the equilibrium price in the later sections.

Remark 3 (Optimal ridge regression). When investor's prior on predictability equals the actual predictability in the predictors, i.e. $\sigma_\beta^2 = ||\beta_1||_2 = \kappa b_*$, the shrinkage τ corresponds to the optimal shrinkage as derived in Kelly et al. (2021). In other words, the estimation accuracy achieves its maximum when the investor knows the correct amount of predictability in his data set versus the amount of predictability still left out. In the following sections I derive results under this assumption, but it will be easy to derive general results for arbitrary σ_β^2 and $\hat{\sigma}^2$. In that case one can just restate results in terms of sufficient statistics $z = \frac{\hat{\sigma}^2}{\sigma_\beta^2}$

and use $\tau = \frac{P_1}{T}z$.

3.3 Risk Premium, Price Volatility and Price Informativeness

In this section I define the objects studied in this paper, including price volatility, risk premium, and price informativeness. I define these objects from the perspective of an econometrician outside the model, who observes an infinite history of payoff f_t and price p_t generated by the investor inside the model.

Specifically, since the model is essentially static, I define the (unconditional) risk premium simply as the expected difference between payoff and price, i.e.

$$RP \equiv E[f_{t+1} - p_t] \tag{15}$$

where the expectation is taken under the true unconditional probability measure. I also define price volatility as the unconditional variance of prices

$$Vol \equiv Var(p_t) \tag{16}$$

Our measure for price informativeness, following Bond et al. (2012) and Bai et al. (2016), is the forecasting price efficiency, which is the variance of the predictable component of payoff f_{t+1} given p_t. Specifically, the econometrician runs an OLS regression of prices on future payoff realization

$$f_{t+1} = \delta_0 + \delta p_t + u_{t+1} \tag{17}$$

and extract the predictive component $\pi_t \equiv \delta_0 + \delta p_t$. Our price informativeness measure is then defined as

$$FPE \equiv Var(f_{t+1}|p_t) = Var(\pi_t) \tag{18}$$

Note that $FPE = Vol$ if the population coefficient $\delta = 1$.

3.4 Benchmark: Sufficient History

In this section I solve for the equilibrium price and price informativeness in a benchmark case, where investor has access to sufficiently long history to estimate $\hat{\beta}$. This means $T \to \infty$ and $P_1/T \to 0$. This scenario corresponds to the traditional asymptotic, where the usual Law of Large Numbers (LLN) applies and the convergence of estimator holds. While these results might seem trivial, they provide a benchmark to which I contrast my main results of learning with insufficient history and data mining.

Lemma 1 (Price with Sufficient History)

When $P_1/T \to 0$, denote $P_x/P = \kappa$, we have

$$\hat{\beta} = (\frac{1}{T}\tilde{X}_T'\tilde{X}_T)^{-1}\frac{1}{T}\tilde{X}_T' f_T \to \begin{pmatrix}\beta_1 \\ 0\end{pmatrix} \tag{19}$$

$$\hat{V}_\beta = (\sigma^2 + b_*(1-\kappa)\sigma_x^2)(\frac{1}{T}\tilde{X}_T'\tilde{X}_T)^{-1}\frac{1}{T^2}\tilde{X}_T'\tilde{X}_T(\frac{1}{T}\tilde{X}_T'\tilde{X}_T)^{-1} \to 0 \tag{20}$$

And as a result, $E_t^I[f_{t+1}|\mathcal{I}_t] = X_{1,t}'\beta_1$ and $V[f_{t+1}|\mathcal{I}_t] = \sigma_x^2 b_*(1-\kappa) + \sigma^2$, and

$$p_t = X_{1,t}'\beta_1 - \rho Q(\sigma_x^2 b_*(1-\kappa) + \sigma^2) \tag{21}$$

Therefore,

$$\begin{aligned} RP &= \rho Q(\sigma_x^2 b_*(1-\kappa) + \sigma^2) \\ Vol = FPE &= \sigma_x^2 b_* \kappa \end{aligned} \tag{22}$$

Furthermore, if the investor observes all predictors $P_x = P$, then we have

$$p_t = X_t'\beta - \rho Q\sigma^2 \tag{23}$$

and

$$\begin{aligned} RP &= \rho Q\sigma^2 \\ Vol = FPE &= \sigma_x^2 b_* \end{aligned} \tag{24}$$

Lemma 1 shows that with sufficient history comparing to the number of parameters, i.e. $P_1/T \to 0$, the investor can recover true β on the truely useful signals and filter out redundant signals by assigning zero coefficients, regardless the distribution of the true parameter and the prior. This is the classic consistency result of Bayesian estimators.

Figure 5 plots price volatility, risk premium, and price informativeness when investor undertakes data mining in a sufficient-history environment. Given no uncertainty in the estimation as $\hat{V}_\beta \to 0$, the risk perceived by the investor comes entirely from the unexplained variation, which comes from the true unlearnable variation σ^2 or unlearnable variation due to unobservable predictors $\sigma_x^2 b_*(1-\kappa)$, and the risk premium is proportional to the summation of these two parts. The price variance, on the other hand, is entirely driven by the variation in the observable predictors, and with correctly estimated κ fraction of true β, the total variation in the prediction is $\kappa b_* \sigma_x^2$. I also note in this case that price is a *conditionally* unbiased forecast for the fundamental, therefore δ in the forecast regression 17 equals 1, and thus $Vol = FEP$. When $\kappa = 1$, the expressions restore to the case equivalent to the investor fully observing the learnable part θ, and all risk and prediction error is driven by the unlearnable variance σ^2. As data mining increases P_1, because $\hat{\beta} = \beta_1$, p_t captures more variation in the payoff more correctly, and therefore the price informativeness increases. This is true even among P_1 predictors some are redundant, because even the investor starts from a nonzero prior on the redundant predictors, she will conclude that their coefficients are zero eventually given sufficient training history.

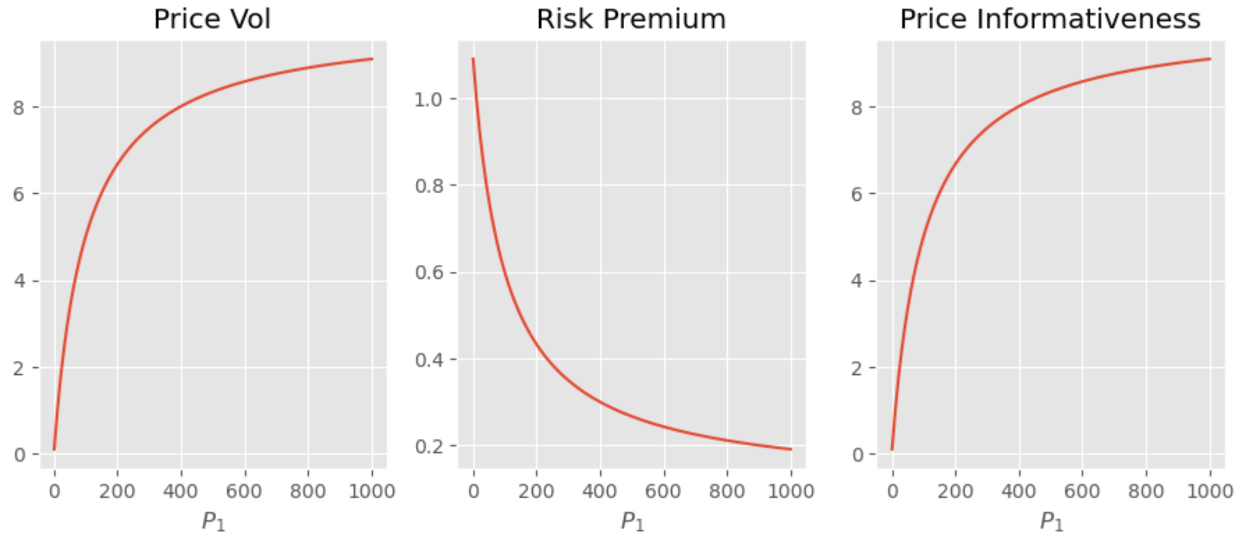


Figure 5: Effect of data mining on Vol, RP and FPE when training data is sufficient $T \to \infty$. I choose $\rho = 1$, $Q = 0.1$, $\sigma^2 = \sigma_x^2 = 1$ and $b_* = 10$. I set true $P = 1000$ and κ to have a polynomial decay as $\kappa(c) = 1 - (1+c)^{-\alpha}$ with $\alpha = 1$.

4 Equilibrium Pricing with Data Mining

In the previous section, Lemma 1 shows that when training data is sufficient, investor can figure out which predictors are useful and estimate their β's correctly, while eliminating redundant predictors by assigning zero coefficients to them. In this section, I move on to study what will happen to asset prices when investor only have finite history to estimate his model. In this regime, training data is scarce relative to signals. As a result, it creates difficulty in investor's learning problem. The main contribution of this paper is to quantify this limits to learning by taking into account the details of the data mining process.

I show that insufficient training data has three effects on investor's parameter estimates. First, it creates a limits to learning, such that investor cannot perfectly estimate $\hat{\beta}$ and estimation uncertainty arises. Second, investor's prior will have nontrivial influence on his estimates. When investor has a non-diffusive prior ($\sigma_\beta^2 < \infty$ in (11)), it can be shown that the explicit shrinkage increases as more predictors are included. Third, in a high-complexity regime where the number of predictors exceeds the number of training data points, an implicit shrinkage arises as there can be multiple parameter solutions to fit the training data, and the model can select the estimate with lowest variance. Combining the three effects from complexity and decreasing return in data efficacy implies asset price patterns distinct from the benchmark case as characterized in Lemma 1.

4.1 Complexity and Data Mining

I introduce the notion of complexity and see how it affects equilibrium prices. I define model complexity as the ratio between P_1 and T, i.e. $c = P_1/T$. Low complexity ($c \approx 0$) describes settings with many more observations than predictors in the model to estimate parameters. This is the regime of traditional econometrics, and the learning result is characterized in Lemma 1. However, in the real world investors only have finite historical observations. Thus, when P_1 increases due to data mining, c starts to deviate from 0. The next proposition characterizes the limiting posterior variance, risk premium, and price informativeness when data mining introduces complexity, using random matrix theory.

Proposition 4.1 (Effect of data mining on equilibrium pricing)
Suppose the investor's information set is described as in Section 3.2. Let $c = P_1/T$, and $\kappa = P_x/P$. As $P_1, T \to \infty$, we have the posterior variance

$$V[\theta|\mathcal{I}_I] \to \mathcal{V}(\tau(c); c, \kappa) \tag{25}$$

the risk premium

$$RP \to \rho Q(\mathcal{V}(\tau(c); c, \kappa) + \sigma^2 + b_* \sigma_x^2 (1-\kappa)) \tag{26}$$

the price volatility

$$Vol \to \mathcal{L}(\tau(c); c, \kappa) \tag{27}$$

and the forecasting price efficiency

$$FPE \to \frac{\mathcal{E}(\tau(c); c, \kappa)^2}{\mathcal{L}(\tau(c); c, \kappa)} \tag{28}$$

where $\tau(c) = c\frac{\sigma^2 + b_* \sigma_x^2(1-\kappa)}{\sigma_\beta^2}$ and

$$\begin{aligned}
\mathcal{E}(\tau(c); c, \kappa) &= b_* \kappa \sigma_x^2 (1 - \tau(c) m(\tau(c); c) \\
\mathcal{V}(\tau(c); c, \kappa) &= (\sigma^2 + b_* \sigma_x^2 (1-\kappa)) \left(c m(\tau(c); c) + c\tau(c) \frac{d}{d\tau} m(\tau(c); c) \right) \\
\mathcal{L}(\tau(c); c, \kappa) &= \mathcal{M}(\tau(c); c, \kappa) + \mathcal{V}(\tau(c); c, \kappa) \\
\mathcal{M}(\tau(c); c, \kappa) &= b_* \kappa \sigma_x^2 (1 - 2\tau(c) m(\tau(c); c) - \tau(c)^2 \frac{d}{d\tau} m(\tau(c); c))
\end{aligned} \tag{29}$$

and $m(\tau;c)$ is the Marcenko-Pastur law

$$m(\tau;c) = \frac{-(1-c+\tau)+\sqrt{(1-c+\tau)^2+4c\tau}}{2c\tau} \tag{30}$$

Proof of Proposition 4.1 relies on characterization of pricing moments when investor uses correctly specified model ($P_x = P$) with complexity. Such characterization also allows us to separate the effect of complexity directly. For brevity I include the discussion of correctly specified model in Appendix A.2.

Proposition 4.1 shows that data mining when training history is insufficient has different implications on asset pricing moments from sufficient history case characterized in Lemma 1. To unpack what drives the changes, it's useful to define quantities when $\sigma_\beta^2 = \infty$, or equivalently $\tau(c) = 0$. It corresponds to results when investor uses a "ridgeless" estimator and when $c < 1$, OLS is the ridgeless estimator. Since investor uses OLS in the benchmark case regardless his prior, studying the ridgeless estimator allows us to separate the effect due to complexity from effect due to investor's prior. The following Lemma unpacks the difference between benchmark and high complexity result.

Lemma 2 (Decompose the complexity wedge)
We have

$$\mathcal{E}(\tau(c);c,\kappa) = \mathcal{E}(0;0,\kappa) - \underbrace{(\mathcal{E}(0;0,\kappa)-\mathcal{E}(0;c,\kappa))}_{\textit{limits to learning}} - \underbrace{(\mathcal{E}(0;c,\kappa)-\mathcal{E}(\tau(c);c,\kappa))}_{\textit{prior bias}} \tag{31}$$

$$\mathcal{V}(\tau(c);c,\kappa) = \underbrace{\mathcal{V}(0;c,\kappa)}_{\textit{ridgeless variance}} - \underbrace{(\mathcal{V}(0;c,\kappa)-\mathcal{V}(\tau(c);c,\kappa))}_{\textit{explicit shrinkage}} \tag{32}$$

$$\mathcal{M}(\tau(c);c,\kappa) = \mathcal{M}(0;0,\kappa) - \underbrace{(\mathcal{M}(0;0,\kappa)-\mathcal{M}(0;c,\kappa))}_{\textit{implicit shrinkage}} - \underbrace{(\mathcal{M}(0;c,\kappa)-\mathcal{M}(\tau(c);c,\kappa))}_{\textit{explicit shrinkage}} \tag{33}$$

where $\mathcal{E}(0;0,\kappa) = \mathcal{M}(0;0,\kappa) = b_*\kappa\sigma_x^2$ is the benchmark price volatility or FPE derived in Lemma 1. Furthermore, we have

$$\mathcal{E}(0;0,\kappa)-\mathcal{E}(0;c,\kappa) = \begin{cases} 0 & \textit{if } c < 1 \\ (1-c^{-1})b_*\kappa\sigma_x^2 & \textit{if } c \geq 1 \end{cases} \tag{34}$$

$$\mathcal{V}(0;c,\kappa) = \begin{cases} \frac{c}{1-c}(\sigma^2+b_*\sigma_x^2(1-\kappa)) & \textit{if } c < 1 \\ \frac{1}{c-1}(\sigma^2+b_*\sigma_x^2(1-\kappa)) & \textit{if } c \geq 1 \end{cases} \tag{35}$$

$$\mathcal{M}(0;0,\kappa)-\mathcal{M}(0;c,\kappa) = \begin{cases} 0 & \textit{if } c < 1 \\ (1-c^{-1})b_*\kappa\sigma_x^2 & \textit{if } c \geq 1 \end{cases} \tag{36}$$

Lemma 2 shows the drivers of pricing moments and derives explicit formula for the difference between benchmark and ridgeless quantities. In the benchmark case, investor's prior doesn't matter and he will run OLS as it's the optimal estimator when there's sufficient data. However, when training data is insufficient, ridgeless estimator will behave differently from OLS with sufficient data, and investor will find it optimal to use a non-zero shrinkage. Both effects will contribute to the deviation from benchmark pricing result. I plot each quantity in Figure 10 in Appendix.

In Propositon 4.1, $\mathcal{E}$ describes the behavior of $Cov(f_{t+1}, p_t)$. In a high complexity world ($P_1 > T$), it's impossible for the investor to construct a conditionally unbiased forecast, even if he starts with non-informative prior. This *limits to learning* channel describes the fundamental challenge in doing estimation when investor faces high dimensional predictors and insufficient data. In addition, when complexity deviates from zero, investor will use a nonzero explicit shrinkage τ if he starts with an informative prior, which will lead to higher conditional bias. This *prior bias* channel further dampens $Cov(f_{t+1}, p_t)$.

In Proposition 4.1, price volatility is driven by two components: the variance in expected price movement from signals (price responsiveness), and estimation uncertainty when investor estimates β. I denote the former by $\mathcal{M}$ and the latter by $\mathcal{V}$. For $\mathcal{V}$, when there's insufficient data, ridgeless variance increases and becomes infinity when $P_1 = T$, where the OLS estimator is undefined. The ridgeless variance declines in the $P_1 > T$ regime, because the multiplicity of least-squares solutions allows ridgeless regression to find a low-norm beta that exactly fits the data. I denote this phenomenon *implicit shrinkage.* The posterior variance is further stabilized as investor chooses a nonzero explicit shrinkage. For $\mathcal{M}$, price responsiveness with ridgeless estimator equals the one for OLS initially, but decreases as $P_1 > T$ for the same reason as *implicit* shrinkage arises. Price responsiveness also declines in explicit shrinkage, as the β estimates are further shrunk towards zero.

The following corollary states that when data efficacy κ has a decreasing return to scale as in Assumption 3, price volatility and informativeness achieves maximum at finite c when investor has uninformative prior. A similar characterization exists for informative prior, and I discuss it in in Appendix A.5.

Corollary 1 (Price volatility and FPE are not always increasing in P_1)
Suppose $\kappa''(c) < 0$, for $\sigma_\beta^2 = \infty$, we have

$$\frac{dVol}{dP_1} < 0; \quad \frac{dFPE}{dP_1} < 0 \tag{37}$$

for large P_1.

Figure 6 shows price volatility, risk premium, and forecasting price efficiency derived in Proposition 4.1, featuring ridgeless and optimal shrinkage. Here I assume κ has polynomial decay, i.e. $\kappa(c) = 1 - (1+c)^{-\alpha}$. I also plot the benchmark results as shown in Figure 5 for better comparison. Panel (a) plots the price volatility. When $P_1 = 0$, the investor doesn't know any signals, and the price doesn't move anywhere. As the investor acquires bigger data set, he finds more true signals and his demand starts to respond to these signals, contributing to an increase in price volatility. When in the world with with insufficient training data, estimation uncertainty starts to show up, pushing up price volatility above the benchmark. Price volatility is undefined for ridgeless investor with $P_1 = T$. Explicit shrinkage helps to mitigate the excess volatility, but price is still more volatile than benchmark. When data mining puts P_1 beyond T, implicit shrinkage emerges and both estimation uncertainty and price responsiveness declines.

Panel (b) in Figure 6 shows the risk premium as a function of P_1. The perceived risk is affected by the estimation uncertainty $V[\theta|\mathcal{I}_I]$ and the unexplanable variation due to missing predictors. The figure shows risk premium with complexity is higher than benchmark. It's because with nonzero complexity, estimation uncertainty cannot be completely eliminated, and this shows up as an additional source of risk, which adds to risk premium.

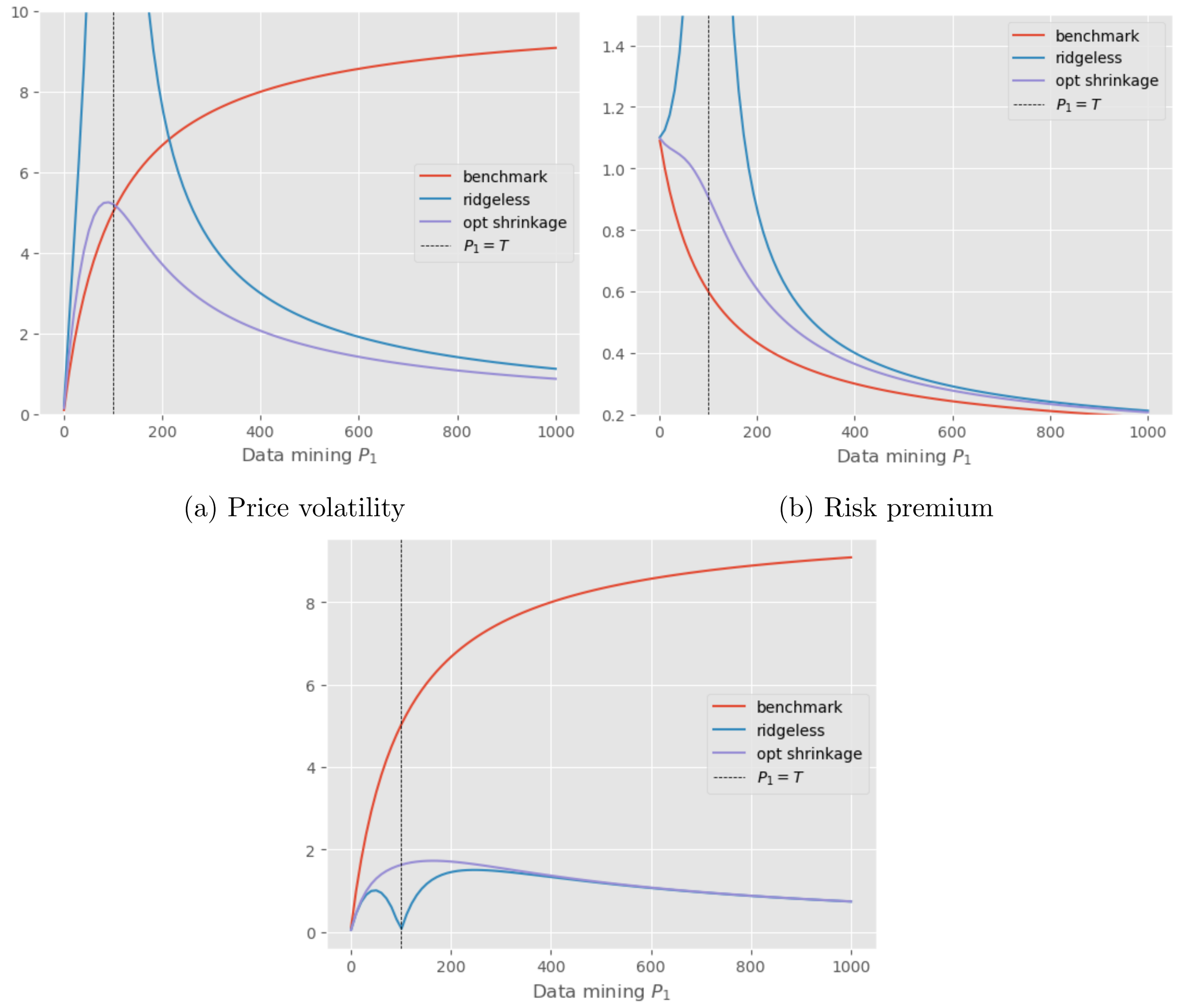
Finally, Panel (d) shows the forecasting price efficiency, whose general pattern is hump-shaped, meaning that price efficiency increases initially but starts to decrease at around $c = 1$. The initial increase is driven by better approximation of the true DGP, but as data efficacy exhibits decreasing return, the gain in better forecasting decreases, while at the same time limits to learning to high complexity kicks in at large P_1 regime. Therefore, price efficiency eventually decreases.

In summary, I have shown that both changes in complexity and data efficacy matters for the equilibrium pricing moments. With decreasing return in data efficacy, data mining can actually lead to lower forecasting power and price efficiency, when the gain in additional true predictors is lower and can't offset the high cost of estimating model in the highly complex regime.

Remark 4 (Changes in optimal shrinkage). I note in Remark 3 the optimal prior (measured from ex-post prediction accuracy) is given by $\sigma_\beta^2 = \kappa b_*$. As data mining increases κ, the optimal shrinkage becomes more diffused to accommodate more predictors. The optimal shrinkage, on the other hand, is given by $\tau(c) = c\frac{\sigma^2 + b_*\sigma_x^2(1-\kappa(c))}{b_*\sigma_x^2\kappa(c)}$. In Appendix A.4 I characterize how optimal shrinkage changes with data mining. For most of the parameter regions the optimal $\tau(c)$ is increasing in data mining scale P_1.

4.2 Value of Data and Optimal Data Mining

In the previous section, I characterized equilibrium price volatility, risk premium, and price informativeness as a function of data mining scale. In this section, I study the change in the investor portfolio's Sharpe ratio due to data mining and characterize the optimal degree of



(a) Price volatility

(b) Risk premium

(c) Forecasting price efficiency (FPE)

Figure 6: Equilibrium pricing result with data mining as a function of P_1 from Proposition A.1. I assume $\rho = 1$, $Q = 0.1$, $\sigma^2 = 1$, $b_* = 10$, and true $P = 10T$ where $T = 100$. "ridgeless" denotes the case when inevstor uses uninformative prior , i.e. $\sigma_\beta^2 = \infty$, while "opt shrinkage" denotes the case when the investor has the correct prior on the variance in β, i.e. $\sigma_\beta^2 = \kappa b_*$. I pick the decay parameter $\alpha = 1$. The dashed line shows the interpolation boundary $c = 1$. The dashed blue line plots the benchmark result as derived in Lemma 1 and shown in Figure 5.

data mining. In this analysis, I assume the representative investor behaves like a price-taker, meaning that data mining only matters in forming informative demand, and the investor doesn't consider the general equilibrium effect of price adjustment due to data mining[11]. This can be rationalized by thinking of the representative investor as a combination of infinite identical investors, each with mass zero, and thus not internalizing the general equilibrium effect. I demonstrate that with sufficient decreasing returns in data efficacy, there's a finite optimal level of data mining level P_1^*, such that the investor's utility will decrease if the data mining scale goes beyond P_1^*. I also show that such data mining level corresponds to the optimal level that maximizes forward price efficiency.

I evaluate Sharpe ratio from the perspective of an outside econometrician, who is able to compute the expectations out-of-sample under the true data generating process. I define the Sharpe Ratio of investor's portfolio as the ratio between expected future wealth divided by its standard deviation, which is proportional to the Sharpe ratio of a market timing strategy using forecast $X_{1,t}\hat{\beta}$ as the position in risky asset:

$$SR \equiv \frac{E_t[W_{t+1}]}{\sqrt{Var_t(W_{t+1})}} \propto \frac{E_t[f_{t+1}X_{1,t}\hat{\beta}]}{Var_t(f_{t+1}X_{1,t}\hat{\beta})} \tag{38}$$

Here investor takes price as given, and it's easy to show that at optimal q_t we have $U(q_t) \propto SR^2$. Therefore, the Sharpe ratio provides a good characterization of the welfare. The next proposition characterizes the (out-of-sample) Sharpe ratio as a function of the data mining process:

Proposition 4.2 (Effect of data mining on Sharpe ratio)
Suppose the investor's information set is described as in Section 3.2. Let $c = P_1/T$, $\kappa = P_x/P_1$. As $P_1, T \to \infty$, we have the squared Sharpe ratio

$$SR^2 \to \frac{\mathcal{E}(\tau(c); c, \kappa)^2}{2\mathcal{E}(\tau(c); c, \kappa)^2 + \mathcal{L}(\tau(c); c, \kappa)(b_*\sigma_x^2 + \sigma^2)} = \frac{1}{2 + (b_*\sigma_x^2 + \sigma^2)FPE^{-1}} \tag{39}$$

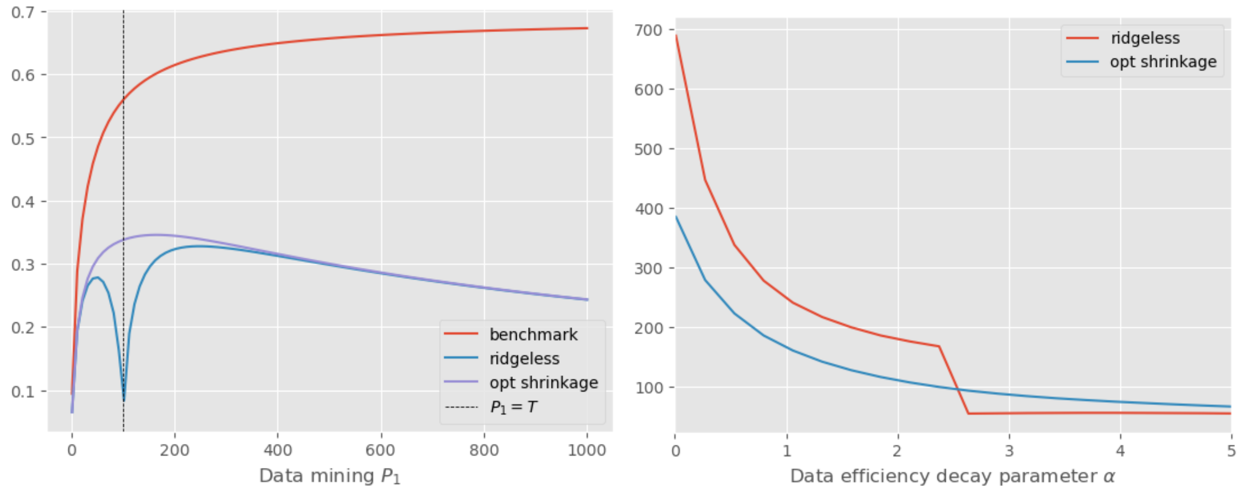
where $\tau(c)$, $\mathcal{E}(\tau(c); c, \kappa)^2$ and $\mathcal{L}(\tau(c); c, \kappa)$ are defined in Proposition 4.1.

Furthermore, when data efficacy κ is a function of c, the optimal data mining is characterized by a c^* such that it solves

$$\max_{c \geq 0} SR(c, \kappa(c)) \tag{40}$$

and is characterized by $\frac{\partial SR(c,\kappa(c))}{\partial c} = 0$.

[11] Because of market clearing, considering the general equilibrium effect will leave the investor's utility unchanged along data mining, because changes in forecast translate one-for-one in prices in the opposite direction, and in equilibrium, data miners will not be better off in any sense. This echoes the Grossman-Stiglitz paradox since there's no exogenous noise in this economy. However, in this case, it becomes hard to rationalize the effort for data mining in the first place, so I proceed by assuming investors are price-takers.



(a) Sharpe ratio

(b) Optimal data mining P_1^*

Figure 7: Sharpe ratio and optimal data mining P_1^* from the perspective of a price-taking representative investor. The left panel shows the Sharpe ratio defined in 38 as a function of data mining level P_1 with different decay parameter $\alpha = 1$, and the right panel shows the Sharpe-maximizing P_1^* as a function of α. Assume $\rho = 1$, $Q = 0.1$, $\sigma^2 = 1$, $b_* = 10$, and true $P = 10T$ where $T = 100$. "ridgeless" denotes the case when inevstor uses uninformative prior , i.e. $\sigma_\beta^2 = \infty$, while "opt shrinkage" denotes the case when the investor has the correct prior on the variance in β, i.e. $\sigma_\beta^2 = \kappa b_*$. The dashed black line shows the interpolation boundary $c = 1$.

Proposition 4.2 characterizes the Sharpe ratio in 38 as a function of c for the price-taking investor. It also shows that there's a one-to-one mapping between the individual investor's Sharpe ratio and the equilibrium forecasting power efficiency. Intuitively, if investor is able to predict payoff at a higher accuracy and derive a higher Sharpe ratio for his portfolio, the equilibrium price – which is generated by the investors – is also more informative about future payoff.

As the FPE exhibits hump-shape with decaying data efficacy, it is as expected that the Sharpe ratio is also hump-shaped. Panel (a) in Figure 7 plots Sharpe ratio as investor increases P_1. In the benchmark case where investor has sufficient training data, obtaining new predictor is always good for investor, because he can estimate parameters precisely and rule out any redundant signals. However, in a complex world, limits to learning and shrinkage means model performance eventually deteriorates as more predictor is discovered. The pattern is similar to FPE in Figure 6.

This implies that there's an optimal P_1^* that maximizes Sharpe ratio and FPE. Further data mining beyond P_1^* exhibits slow increase in the approximation capacity that doesn't justify the increase in limits to learning imposed by higher complexity. Panel (b) in Figure 7 shows the optimal P_1^* as a function of how fast data efficacy decays. When α is high (lower), the decay is faster (slower), thus the investor should stop data mining sooner (later), meaning a lower (higher) P_1^*.

Remark 5 (Comparing to the "Virtue of Complexity"). It is important to compare my results with the recent finding of "virtue of complexity" in asset pricing, as discovered in Kelly et al. (2021) and Didisheim et al. (2023). They also study empirical models whose parameterization is large relative to training data and find that complex models outperform simple ones. One key distinction in my setup versus theirs is the assumption on how the model size expands. In Kelly et al. (2021) and Didisheim et al. (2023), the empirical model gets bigger because of an increase in features that carry the same predictive content to the payoff. In other words, there is a *constant* return to scale. Such assumption applies to a modeler who is given a fixed set of raw predictors, doesn't know the true functional form of the predictive relationship, but instead is motivated by universal approximation theorem (e.g. Hornik (1991)) and uses a basic neural network to approximate the functional form. An expansion of model complexity can thus be thought of an increase in the number of hidden nodes in the neural network or an increase in nonlinear basis functions. In this paper, I complement their finds by considering what will happen when the raw predictor set increases, while taking the functional form fixed (e.g. linear predictive relationship). As shown in Figure 3, increasing predictors doesn't always lead to higher performance once the neural network size is fixed. The decreasing returns in data efficacy is a realistic assumption that captures the challenge in mining alternative data and a key driver of my main result.

Remark 6 (Difficulty in finding the optimum). If the functional form of κ is known, the investor will stop mining additional data sets once the size of data set reaches the optimal P_1. However, in the real world, investors don't know the true data efficacy, and they can only do so by evaluating their model out-of-sample. Given a finite out-of-sample period, it will also be challenging for investors to draw conclusion on when to stop data mining. Additionally, investor might engage in excess data mining if they have other incentives aside from Sharpe maximization. For example, if it's desirable for a hedge fund to provide distinct strategies compared to their peers, as it allows the fund manager to attract more asset under management, the hedge fund might want to use more alternative data that is unexplored by others, even if the data efficacy might be low. For these reasons, in this paper I mainly focus on the comparative statics of data mining, while the empirical evaluation on optimal data mining is left for future research.

5 Empirical Results

In this section, I examine the empirical impact of increasing alternative data on price informativeness. By leveraging the release of satellite coverage data for retailer stocks as an exogenous expansion of the investor's potential signal set, I employ a difference-in-difference approach to identify the causal effect of this expansion on price informativeness. The findings reveal that the release of satellite data in fact reduces price informativeness of the covered, with a more pronounced effect observed for firms with shorter data availability (higher complexity)[12]. These results support the idea that data mining can be detrimental to price efficiency and underscore the role of training history availability in shaping price efficiency in the real world.

5.1 Data

The main data set comes from RS Metrics, the first data vendor to introduce real-time parking lot traffic signals derived from satellite images in the US. This data set comprises daily store-level information about parking lot capacity and utilization for major public retailers, such as Walmart, Bed Bath & Beyond, Target etc. RS Metrics processes raw satellite images to generate parking lot traffic signals, which are then sold to institutional investors who might be utilizing these signals in their trading strategies.

I obtain the start dates of traffic signal coverage from RS Metrics, which starts in Q2:2009. Different stocks will have different coverage start dates, which allows me to investigate the effect of data release as well as training history length on price informativeness. I obtain the release dates of traffic signals to investors from Katona et al. (2022), which is usually one year after RS Metrics start covering the stock. In total there are 52 retailer stocks within the coverage. In addition, we obtain market-based information such as market cap, volume and return from CRSP, and accounting information such as total asset, book value, and earnings from Compustat. I use *Operating Income After Depreciation* (QIADPQ) as our measure of earnings at the quarterly frequency. My sample starts in Q1:2003 and ends in Q4:2020 to ensure sufficient observations pre- and post- satellite data release, allowing for a more accurate measurement of price informativeness.

5.2 Empirical Design

To perform the difference-in-difference analysis, I construct the control group by selecting firms with the same 6-digit Global Industry Classification Standard (GICS) codes as the covered firms that don't have satellite data coverage. For each covered firm, I then match

[12] I note that this result contrasts Zhu (2019) who finds an increase in price discovery with the release of satellite data, and Katona et al. (2022) who reports no significant impact on price discovery. The main difference between my empirical approach and previous literature is that both Zhu (2019) and Katona et al. (2022) measure price discovery using abnormal returns around earnings announcement dates, whereas I follow the standard literature on measuring price informativeness as in Bai et al. (2016); Farboodi et al. (2022a) and measure FPE by regressing future earnings on current prices at a quarterly basis. In addition, as Katona et al. (2022) points out, Zhu (2019) blends firms that are truly covered by satellite data with firms in the same industry but not covered, and Zhu (2019) assumes the release dates that might be different from the actual release dates provided by RS Metrics.

it with three control firms with most similar book-to-market ratio, gross profitability, sales-to-equity ratio, and return on equity. This matching is achieved by minimizing the total distance in these characteristics across matched pairs. Importantly, my construction avoids using early treated firms as control firms for late treated ones, and thus my estimates are not impacted by comparing early-treated groups to late treated-groups that typically confound staggered Diff-in-Diff estimates. After removing firms with incomplete stock-level data, the final sample comprises 50 treated firms and 93 control firms.

I first measure price informativeness in the similar spirit as Bai et al. (2016); Farboodi et al. (2022a). Specifically, for each stock i at quarter t, I run a stock-level time series regression

$$\frac{E_{i,t+\tau+1}}{A_{i,t+\tau}} = \alpha_{i,t} + \beta_{i,t} \cdot \frac{M_{i,t+\tau}}{A_{i,t+\tau}} + \gamma_{i,t} X_{i,t} + \epsilon_{i,t} \tag{41}$$

for $\tau = 0, 1, \cdots 20$ quarters, where $E_{i,t+\tau+1}/A_{i,t+\tau}$ is the cash-flow (or more precisely, operating income after depreciation) of firm i in quarter $t+\tau+1$ scaled by its total asset in quarter $t+\tau$; $M_{i,t+\tau}/A_{i,t+\tau}$ is market capitalization scaled by total asset, and $X_{i,t}$ is the controls, which includes past earnings to asset. This is the analog to the payoff prediction regression as in 17. I then compute the FPE estimate as

$$F\hat{P}E_{i,t} = \hat{\beta}_{i,t} \cdot \sigma_{i,t}^{M/A} \tag{42}$$

where $\sigma_{i,t}^{M/A}$ is the standard deviation of market-to-asset ratio used in the regression. The scaling follows Bai et al. (2016); Farboodi et al. (2022a) and my final $F\hat{P}E_{i,t}$ can be viewed as the empirical counterpart to 18.

I then estimate the impact of satellite data release on price informativeness in a difference-in-difference design. Specifically, we run the following regression

$$F\hat{P}E_{i,t} = \beta_1 Post_{i,t} + \beta_2 Treat_{i,t} + \beta_3 Post_{i,t} \times Treat_{i,t} + Controls_{i,t} + \epsilon_{i,t} \tag{43}$$

where $Post_{i,t}$ is an indicator if the current t is after satellite data release for firm i, and $Treat_{i,t}$ is an indicator if the firm is in the treatment group. I include controls that might affect price discovery, including firm size (log market cap), quarterly trading volume, and quarterly return. The estimate of interest is β_3, which estimates the causal impact of satellite release on forward price efficiency.

In addition, the model implies that signals with shorter history (higher complexity) might lead to lower price informativeness. To test this implication, I run the following regression

$$\begin{aligned} F\hat{P}E_{i,t} = {} & \beta_1 Post_{i,t} + \beta_2 Treat_{i,t} + \beta_3 Post_{i,t} \times Treat_{i,t} \\ & + \beta_4 Complexity_{i,t} \times Treat_{i,t} + Controls_{i,t} + \epsilon_{i,t} \end{aligned} \tag{44}$$

where I define $Complexity_{i,t}$ as the inverse of the number of quarters since RS Metrics start covering the firm (in other words $1/T$ in the theoretical model). The coefficient β_4 estimates the effect of complexity on price informativeness for the treated firms.

5.3 Result

Table 1 presents the difference-in-difference coefficient estimates for the price informativeness measure using one-quarter ahead earnings forecast regression. The significantly negative coefficient on $Post \times Treat$ indicates that the release of satellite data leads to lower price informativeness. This surprising result is contrary to the common view that the discovery of alternative data should enhance price informativeness, but is in line with my model prediction. In the second column, I report the impact of complexity on price informativeness. The significantly negative coefficient on $Complexity \times Treat$ suggests that firms with signals of higher estimation complexity (shorter history available) experience a more pronounced decline in price informativeness. This result is consistent with my model that demonstrates that estimation complexity hinders price discovery.

In addition to measuring price informativeness using one-quarter ahead forecast, I also conduct the analysis on price informativeness measured with longer forecast horizon. Table 2 reports the difference-in-difference coefficient estimates for price informativeness measure using 2,3,4-quarter ahead earnings forecast regression. I again find significant reduction in long-term price informativeness after the release of satellite data. The effect on complexity however becomes less significant at longer horizon. I note that my analysis only speaks to the effect of a particular alternative data set for a particular set of stocks, and a comprehensive analysis would require researcher to know all the data sets used by investors, which is very challenging in the real world. Nonetheless, these empirical results illustrate that the discovery of new signals might be detrimental to price informativeness, a phenomenon my model provides a potential explanation for.

	(1)	(2)
$Post \times Treat$	-0.006***	-0.005***
	(-6.49)	(-4.54)
$Complexity \times Treat$		-0.0096**
		(-1.97)
$Post$	0.008***	0.009***
	(9.11)	(9.22)
$Treat$	0.004***	0.004***
	(9.22)	(9.32)
Controls	Yes	Yes
Time FE	Yes	Yes
R^2	0.0556	0.0572
Obs	2531	2531

Table 1: This table shows that satellite image data release decreases forward price efficiency, and the decrease is more significant for stocks with higher complexity. I report the estimates based on DID regression model in equation (43) and (44). The dependent variable $F\hat{P}E$ is estimated from a rolling regression with 20 quarters of one-quarter ahead earnings on current market cap, as in (41). $Post$ is an indicator that takes the value one after the release of RS Metrics coverage. $Treat$ is an indicator variable that takes the value one for the treated group of retailers with satellite coverage. The controls include log market capitalization, quarterly dollar trading volume, and quarterly return.***, **, and * indicate statistical significance at 1%, 5% and 10% level, based on two-tailed tests.

	2Q ahead	2Q ahead	3Q ahead	3Q ahead	4Q ahead	4Q ahead
$Post \times Treat$	-0.005***	-0.003***	-0.003***	-0.003***	-0.004***	-0.005***
	(-4.66)	(-2.79)	(-3.34)	(-3.32)	(-5.10)	(-5.66)
$Complexity \times Treat$		-0.012**		0.004		0.002
		(-2.31)		(0.089)		(1.28)
$Post$	0.008***	0.008***	0.004***	0.004***	0.004***	0.004***
	(8.25)	(8.37)	(5.71)	(5.64)	(5.63)	(5.48)
$Treat$	0.003***	0.003***	-0.001***	-0.001***	-0.002***	-0.002***
	(6.47)	(6.59)	(-2.16)	(-2.18)	(-6.27)	(-6.40)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.0432	0.0453	0.0245	0.0248	0.0441	0.0465
Obs	2531	2531	2527	2527	2514	2514

Table 2: This table shows that satellite image data release decreases forward price efficiency. I report the estimates based on DID regression model in equation (43) and (44). The dependent variable $F\hat{P}E$ is estimated from a rolling regression with 20 quarters of 2-quarter, 3-quarter, and 4-quarter ahead earnings on current market cap, as in (41). $Post$ is an indicator that takes the value one after the release of RS Metrics coverage. $Treat$ is an indicator variable that takes the value one for the treated group of retailers with satellite coverage. The controls include log market capitalization, quarterly dollar trading volume, and quarterly return.***, **, and * indicate statistical significance at 1%, 5% and 10% level, based on two-tailed tests.

6 Conclusion

In this paper, I demonstrate that data mining can have a dual impact on information efficiency. While data mining discovers more true predictors, expanding the set of predictive signals increases the challenge of estimation due to insufficient training history. Unlike other machine learning and AI applications such as auto-driving cars and natural language processing, where training data is abundant and easy to generate, in the real world investors can't easily generate time series return data. This leads to model complexity and limits to learning. I theoretically analyze how price volatility, risk premium, and price informativeness respond to the escalation of data mining efforts. In conjunction with a decreasing return in data efficacy, I show that there exists a finite optimal level of data mining scale. Acquiring further predictive signals beyond the optimum would lead to lower price informativeness and Sharpe ratio. As a result, investors would optimally choose to ignore some costless signals, even they contain information that are useful for prediction. These theoretical insights contribute to understanding asset pricing in high-dimensional settings, while future research could explore empirical approaches targeting the high-dimensionality of investors' estimation problems.

References

Al-Najjar, N. I. (2009). Decision makers as statisticians: Diversity, ambiguity, and learning. *Econometrica*, 77(5):1371–1401.

Aragones, E., Gilboa, I., Postlewaite, A., and Schmeidler, D. (2005). Fact-free learning. *American Economic Review*, 95(5):1355–1368.

Bai, J., Philippon, T., and Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics*, 122(3):625–654.

Bai, Z. and Zhou, W. (2008). Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442.

Balasubramanian, A. and Yang, Y. D. (2019). Trading with high-dimensional data. *Available at SSRN 3583217*.

Banerjee, S., Davis, J., and Gondhi, N. (2018). When transparency improves, must prices reflect fundamentals better? *The Review of Financial Studies*, 31(6):2377–2414.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics*, 22(1):107–111.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bond, P., Edmans, A., and Goldstein, I. (2012). The real effects of financial markets. *Annu. Rev. Financ. Econ.*, 4(1):339–360.

Burkholder, D. L. (1966). Martingale transforms. *The Annals of Mathematical Statistics*, 37(6):1494–1504.

Chen, A. Y., Lopez-Lira, A., and Zimmermann, T. (2022). Peer-reviewed theory does not help predict the cross-section of stock returns. *arXiv preprint arXiv:2212.10317*.

Collin-Dufresne, P., Johannes, M., and Lochstoer, L. A. (2016). Parameter learning in general equilibrium: The asset pricing implications. *American Economic Review*, 106(3):664–698.

Dávila, E. and Parlatore, C. (2023). Identifying price informativeness. Technical report, National Bureau of Economic Research.

Didisheim, A., Ke, S., Kelly, B. T., and Malamud, S. (2023). Complexity in factor pricing models. *Swiss Finance Institute Research Paper*, (23-19).

Dugast, J. and Foucault, T. (2018). Data abundance and asset price informativeness. *Journal of Financial economics*, 130(2):367–391.

Dugast, J. and Foucault, T. (2023). Equilibrium data mining and data abundance. *HEC Paris Research Paper No. FIN-2020-1393, Université Paris-Dauphine Research Paper*, (3710495).

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Farboodi, M., Matray, A., Veldkamp, L., and Venkateswaran, V. (2022a). Where has all the data gone? *The Review of Financial Studies*, 35(7):3101–3138.

Farboodi, M., Singal, D., Veldkamp, L., and Venkateswaran, V. (2022b). Valuing financial data. Technical report, National Bureau of Economic Research.

Fernández-Villaverde, J. (2021). Has machine learning rendered simple rules obsolete? *European Journal of Law and Economics*, pages 1–15.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, 33(5):2326–2377.

Ghosh, N. and Belkin, M. (2022). A universal trade-off between the model size, test loss, and training loss of linear predictors. *arXiv preprint arXiv:2207.11621*.

Goldstein, I., Spatt, C. S., and Ye, M. (2021). Big data in finance. *The Review of Financial Studies*, 34(7):3213–3225.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance*, 72(4):1399–1440.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2022). Is there a replication crisis in finance? *The Journal of Finance.*

Katona, Z., Painter, M., Patatoukas, P., and Zeng, J. (2022). On the capital market consequences of alternative data: Evidence from outer space. *Available at SSRN 3222741.*

Kelly, B., Malamud, S., and Zhou, K. (2021). The virtue of complexity in return prediction. *Swiss Finance Institute Research Paper*, (21-90).

Kelly, B. T. and Xiu, D. (2023). Financial machine learning. *Available at SSRN.*

Lewellen, J. and Shanken, J. (2002). Learning, asset-pricing tests, and market efficiency. *The Journal of finance*, 57(3):1113–1145.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 34(1):1–18.

Martin, I. W. and Nagel, S. (2022). Market efficiency in the age of big data. *Journal of financial economics*, 145(1):154–177.

Montiel Olea, J. L., Ortoleva, P., Pai, M. M., and Prat, A. (2022). Competing models. *The Quarterly Journal of Economics*, 137(4):2419–2457.

Nagel, S. (2021). *Machine learning in asset pricing*, volume 8. Princeton University Press.

Pastor, L. and Veronesi, P. (2009). Learning in financial markets. *Annu. Rev. Financ. Econ.*, 1(1):361–381.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

Silverstein, J. W. and Bai, Z. (1995). On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192.

Veldkamp, L. (2023a). Valuing data as an asset. *Review of Finance*, page rfac073.

Veldkamp, L. L. (2023b). *Information choice in macroeconomics and finance.* Princeton University Press.

Vives, X. (2010). *Information and learning in markets: the impact of market microstructure.* Princeton University Press.

Zhang, L. (2005). The value premium. *The Journal of Finance*, 60(1):67–103.

Zhu, C. (2019). Big data as a governance mechanism. *The Review of Financial Studies*, 32(5):2021–2061.

A Additional Theoretical Results

A.1 Decreasing returns to scale from decreasing predictability

In Assumption 1 and the formulation of data mining in Section 3.2, I assume predictive coefficient β is ex-ante the same for all predictors, and the decreasing return in data efficacy comes from inclusion of redundant signals. Here I present another formulation of decreasing return in data efficacy, arising from a decreasing predictability from newly discovered signals.

I continue to assume that all true predictors $\tilde{X}_{i,t}$ are i.i.d with $E[\tilde{X}_{i,t}] = 0$ and $E[\tilde{X}_{i,t}^2] = \sigma_x^2$. However, instead of assuming $E[\beta_i^2] = b_*/P$ for all i, I relax this assumption by denoting $E[\beta_i^2] = b_*^i/P$, where a higher b_*^i corresponds to a higher predictability associated with predictor i. I assume there's no redundant signals, so that all signals discovered by the data miner are true. In this case, one can define data efficacy as the fraction of predictability covered in the predictor set the data miner owns:

$$\kappa = \frac{\sum_{i=1}^{P_1} b_*^i}{\sum_{i=1}^{P} b_*^i} \tag{45}$$

A decreasing return in data efficacy then means signals have decreasing predictability in the order of discovery, i.e.

$$b_*^1 > b_*^2 > b_*^3 \cdots > b_*^{P_1} \tag{46}$$

The natural interpretation of this formulation is that investor first discover the most salient and prominent signals, and once those strong predictors are discovered, investors then move on to discover weak signals. Figure 8 illustrates this data mining process. I note that all of our theory goes through when we denote $b_* = \sum_i^P b_*^i$.

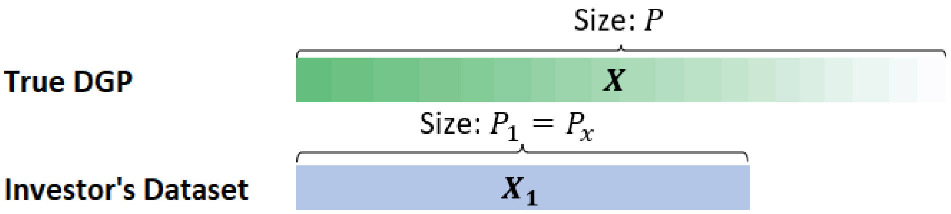


Figure 8: Alternative data mining process, where investor discovers less important signals when they undertake data mining.

A.2 Correctly specified model

Proposition A.1 (Complexity in correctly specified model)
Suppose the model is correctly specified such that $P_x = P_1 = P$. Let $c = P_1/T$ as $P_1, T \to \infty$.

We have the posterior variance

$$V[\theta|\mathcal{I}_I] \to \mathcal{V}(\tau(c);c) \tag{47}$$

the risk premium

$$RP \to \rho Q(\mathcal{V}(\tau(c);c) + \sigma^2) \tag{48}$$

the price volatility

$$Vol \to \mathcal{L}(\tau(c),c) \tag{49}$$

and the forecasting price efficiency

$$FPE \to \frac{\mathcal{E}(\tau(c);c)^2}{\mathcal{L}(\tau(c);c)} \tag{50}$$

where $\tau(c) = c\frac{\sigma^2}{\sigma_\beta^2}$ *and*

$$\mathcal{V}(\tau(c);c) = c\sigma^2\sigma_x^2 m(\tau(c);c) + c\tau(c)\sigma^2\sigma_x^2\frac{d}{d\tau}m(\tau(c);c) \tag{51}$$

$$\mathcal{E}(\tau(c);c) = b_*\sigma_x^2(1-\tau(c)m(\tau(c);c) \tag{52}$$

$$\mathcal{L}(\tau(c);c) = b_*\sigma_x^2(1-2\tau(c)m(\tau(c);c) - \tau(c)^2\frac{d}{d\tau}m(\tau(c);c)) + \mathcal{V}(\tau(c);c) \tag{53}$$

and $m(\tau;c)$ *is the Marcenko-Pastur law*

$$m(\tau;c) = \frac{-(1-c+\tau) + \sqrt{(1-c+\tau)^2 + 4c\tau}}{2c\tau} \tag{54}$$

To derive the expression for misspecifed case, notice that with isotropic assumption of X and the rotational symmetry assumption of β in Assumption 1, we can effectively write the data generating process of f_{t+1} as

$$f_{t+1} = \beta_1' X_{1,t} + \underbrace{\beta_2' X_{2,t} + \epsilon_{t+1}}_{u_{t+1}} \tag{55}$$

where u_{t+1} is the unpredictable term from investor's view, which includes both the truly unpredictable component ϵ_{t+1} and the unpredictable component due to unobservable predictors $\beta_2' X_{2,t}$. Therefore, u_{t+1} is independent of $X_{1,t}$ and W_t, having mean zero and variance $\sigma^2 + ||\beta_2||_2^2\sigma_x^2$. Thus, the risk premium, out-of-sample expected R^2, and the forecasting price efficiency will behave exactly the same as we computed in Proposition A.1, after we make the substitution with

$$b_* \mapsto b_*\kappa \quad \text{and} \quad \sigma^2 \mapsto \sigma^2 + b_*\sigma_x^2(1-\kappa)$$

and putting these observations together leads to Proposition 4.1.

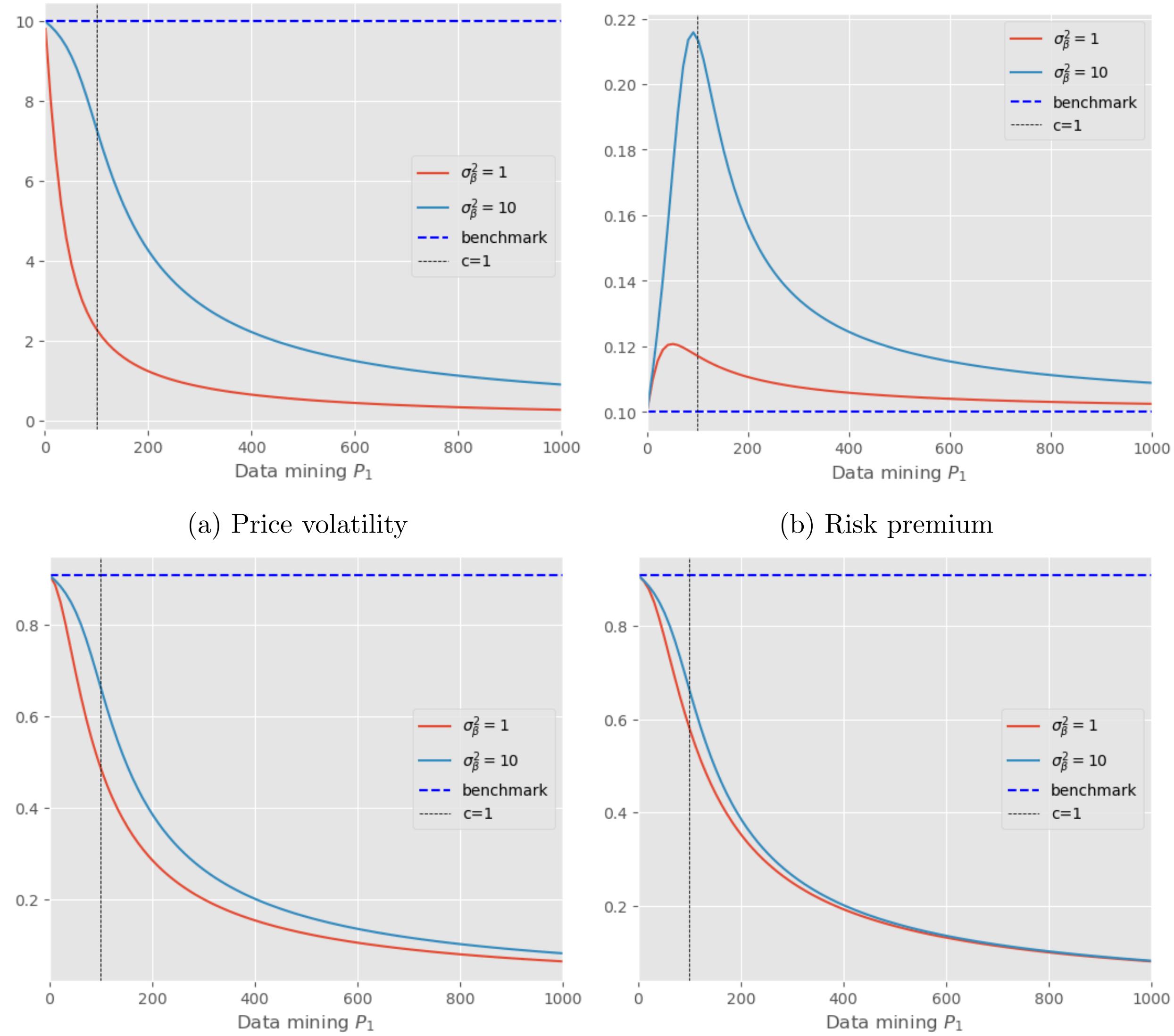
Proposition A.1 demonstrates how insufficient training history ($c > 0$) affects equilibrium risk premium, volatility and price informativeness. Note here all results are driven by changes in complexity (equivalently change in P_1 when T is taken as fixed). Figure 9 plots these theoretical results as P_1 increases. Panel (a) shows that price volatility decreases as P_1 gets large. This is driven by increase in both the explicit shrinkage investor uses in estimation, i.e. $\tau(c) = c\frac{\sigma^2}{\sigma_\beta^2}$[13], and implicit shrinkage when $c > 1$. For implicit shrinkage, intuitively, as there are more predictors than the observations, investor will find it easier to have linear combinations of predictors to fit training data well. Given that the investor is Bayesian with non-diffusive prior, he will settle on a $\hat{\beta}$ with smaller norm (and smaller posterior variance). This is usually referred to as the "benign overfit" or "virtue of complexity" phenomenon and is actively being explored by machine learning researchers. (e.g. Hastie et al. (2022), Kelly et al. (2021), Didisheim et al. (2023)).

When P_1 is close to 0, there's no uncertainty in estimating β and price fully reflects variations in X_t, which means the variance in prices equals to $b_*\sigma_x^2$ in expectation. However, when P_1 becomes large, both explicit and implicit shrinkage penalizes $\hat{\beta}$ towards 0. Therefore, it reduces the variance in the forecast, which means price volatility decreases.

Panel (b) plots the risk premium as a function of P_1. When P_1 is close to zero, correct specification of the model and consistent estimator imply all explainable variation in the payoff is captured. Therefore, risk premium is only related to the unlearnable variation σ^2 (the benchmark case). However, as P_1 starts to increase, estimation uncertainty about β kicks in, which serves as an additional source of risk and increases risk premium. As data mining further increases P_1 beyond $c = 1$, both explicit and implicit shrinkage increase and penalizes parameter estimates towards zero. Thus, the "subjective" uncertainty decreases from the perspective of investor, and so is the risk premium.

Panel (c) plots the out-of-sample R^2 when treating price as the forecast for fundamental, and panel (d) plots the forecasting price efficiency. We see that when P_1, both R^2 and FPE equal to the fraction of learnable variation in the payoff, as shown in Lemma 1. However, when P_1 increases, the increasing explicit shrinkage induce a larger bias in the forecast, resulting in an initial decrease in the $c < 1$ region. As P_1 increases further into the $c > 1$ region, the implicit shrinkage also arises and contributes to a further decrease in the variance in $\hat{\beta}$. As a result, price becomes less sensitive to changes in the predictors and price informativeness decreases.

[13] Note that $\tau(c)$ corresponds to the optimal shrinkage that attains the highest Sharpe ratio and out-of-sample R^2 derived in Kelly et al. (2021).



(a) Price volatility

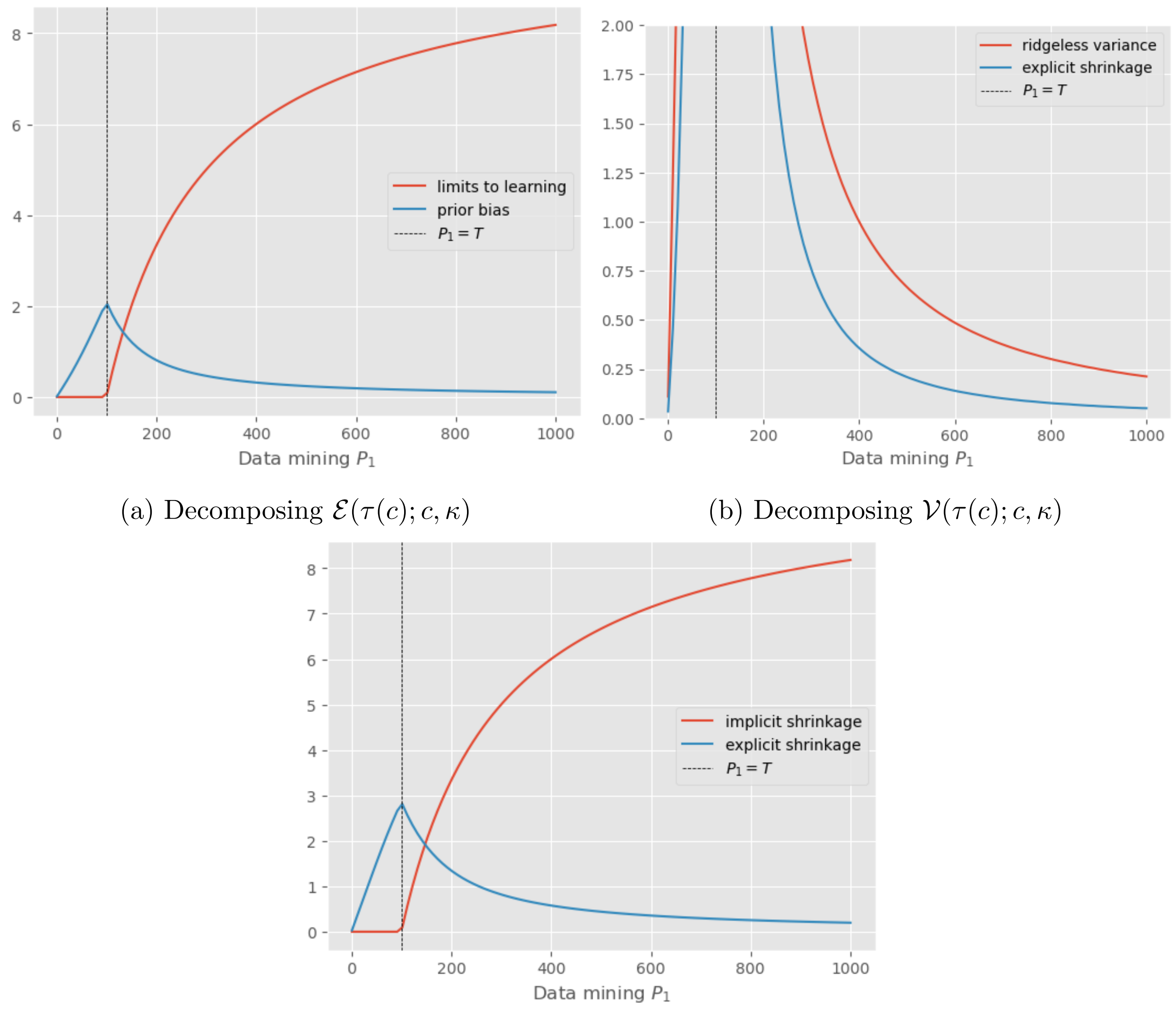
(b) Risk premium

(c) Out-of-sample R^2

(d) Forecasting price efficiency (FPE)

Figure 9: Equilibrium pricing result with complexity in correctly specified model as a function of P_1 and σ_β^2 from Proposition A.1. We assume $\rho = 1$, $Q = 0.1$, $\sigma^2 = 1$, $b_* = 10$, and true $P = 10T$ where we fix $T = 100$. The dashed black line shows the interpolation boundary $c = 1$. The dashed blue line plots the benchmark result as derived in Lemma 1.

A.3 Decomposing the complexity wedge



(a) Decomposing $\mathcal{E}(\tau(c); c, \kappa)$

(b) Decomposing $\mathcal{V}(\tau(c); c, \kappa)$

(c) Decomposing $\mathcal{M}(\tau(c); c, \kappa)$

Figure 10: Decomposing the complexity wedge following quantities in Lemma 2. I assume $\rho = 1$, $Q = 0.1$, $\sigma^2 = 1$, $b_* = 10$, and true $P = 10T$ where we fix $T = 100$. The dashed black line shows the interpolation boundary $c = 1$.

A.4 Changes in optimal shrinkage

In Remark 4 we show that the optimal shrinkage is given by

$$\tau(c) = c\frac{\sigma^2 + b_*\sigma_x^2(1 - \kappa(c))}{b_*\sigma_x^2\kappa(c)} \tag{56}$$

and thus

$$\frac{\partial\tau(c)}{\partial c} = \frac{1}{b_*\sigma_x^2\kappa^2(c)}\left[(\sigma^2 + b_*\sigma_x^2(1 - \kappa(c)))(\kappa(c) - c\kappa'(c)) - b_*\sigma_x^2 c\kappa'(c)\kappa(c)\right] \tag{57}$$

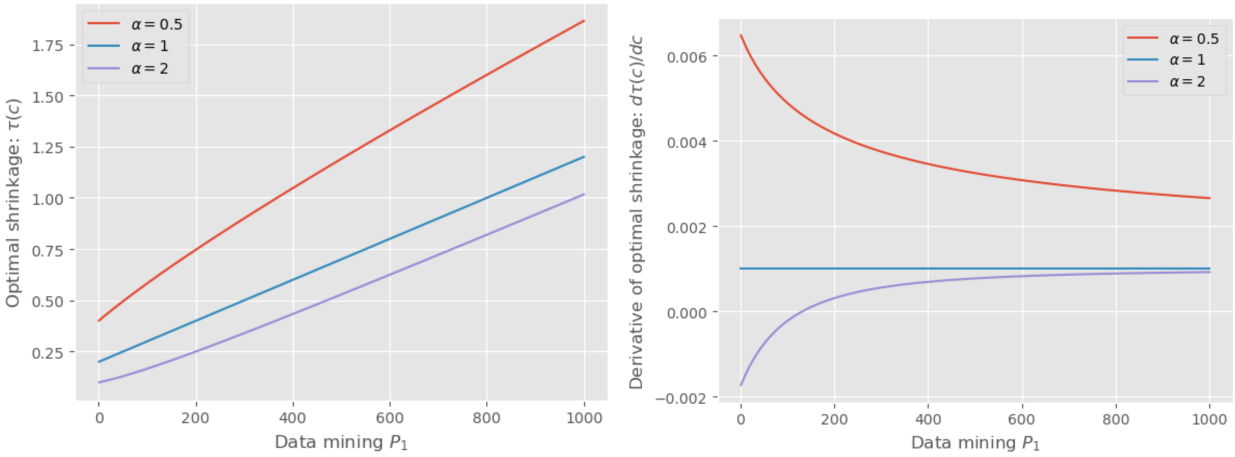
Given $\kappa''(c) < 0$, we know that $\kappa(c) - c\kappa'(c) > 0$, and the second term

$$b_* \sigma_x^2 c\kappa'(c)\kappa(c) \leq b_* \sigma_x^2 \kappa^2(c) \tag{58}$$

Thus, the sufficient condition for $\frac{\partial \tau(c)}{\partial c} > 0$ is that

$$\sigma^2 + b_* \sigma_x^2 (1 - \kappa(c)))(\kappa(c) - c\kappa'(c) \geq b_* \sigma_x^2 \kappa^2(c) \tag{59}$$

Figure 11 shows the optimal shrinkage as a function of data mining P_1 for polynomial decaying κ. We see that for most specifications of α, the optimal shrinkage is increasing in P_1.



(a) Optimal shrinkage $\tau(c)$ as a function of P_1

(b) Derivative of optimal shrinkage $\tau(c)$

Figure 11: Optimal shrinkage and its derivative to P_1 for κ following polynomial decay with different decay parameters.

A.5 Comparative statics with optimal shrinkage

From direct calculation we have

$$\begin{aligned} \frac{d\mathcal{E}(\tau(c); c, \kappa)}{dc} &= \frac{\partial \mathcal{E}}{\partial c} + \frac{\partial \mathcal{E}}{\partial \tau}\frac{d\tau}{dc} + \frac{\partial \mathcal{E}}{\partial \kappa}\kappa'(c) \\ &= b_* \sigma_x^2 \left[\kappa \left(1 - \tau \frac{\partial m}{\partial c} - \left(m + \tau \frac{\partial m}{\partial \tau} \right) \right) \frac{d\tau}{dc} + (1 - \tau m(\tau, c))\kappa'(c) \right] \end{aligned} \tag{60}$$

$$\begin{aligned}\frac{d\mathcal{M}(\tau(c);c,\kappa)}{dc} &= \frac{\partial\mathcal{M}}{\partial c} + \frac{\partial\mathcal{M}}{\partial\tau}\frac{d\tau}{dc} + \frac{\partial\mathcal{M}}{\partial\kappa}\kappa'(c)\\ &= b_*\sigma_x^2\Bigg[\kappa\left(1 - 2\tau\frac{\partial m}{\partial c} - \tau^2\frac{\partial^2 m}{\partial\tau\partial c} - \left(2m + 4\tau\frac{\partial m}{\partial\tau} + \frac{\partial^2 m}{\partial\tau^2}\right)\frac{d\tau}{dc}\right)\\ &+ (1 - 2\tau m(\tau,c) - \tau^2\frac{\partial m}{\partial\tau})\kappa'(c)\Bigg]\end{aligned} \tag{61}$$

and

$$\begin{aligned}\frac{d\mathcal{V}(\tau(c);c,\kappa)}{dc} &= \frac{\partial\mathcal{V}}{\partial c} + \frac{\partial\mathcal{V}}{\partial\tau}\frac{d\tau}{dc} + \frac{\partial\mathcal{V}}{\partial\kappa}\kappa'(c)\\ &= (\sigma^2 + b_*\sigma_x^2(1-\kappa))\Bigg[\left(m + c\frac{\partial m}{\partial c} + \tau\frac{\partial m}{\partial\tau} + c\tau\frac{\partial^2 m}{\partial\tau\partial c}\right)\\ &+ \left(2c\frac{\partial m}{\partial\tau} + c\tau\frac{\partial^2 m}{\partial\tau^2}\right)\frac{d\tau}{dc}\Bigg]\\ &- b_*\sigma_x\left(cm + c\tau\frac{\partial m}{\partial\tau}\right)\kappa'(c)\end{aligned} \tag{62}$$

and these quantities can be used to characterize

$$\frac{dVol}{dc} = \frac{d\mathcal{M}(\tau(c);c,\kappa)}{dc} + \frac{d\mathcal{V}(\tau(c);c,\kappa)}{dc} \tag{63}$$

and

$$\frac{dFPE}{dc} = \frac{2\mathcal{E}(\tau(c);c,\kappa)\frac{d\mathcal{E}(\tau(c);c,\kappa)}{dc}Vol - \mathcal{E}(\tau(c);c,\kappa)^2\frac{dVol}{dc}}{Vol^2} \tag{64}$$

We don't have a closed form characterization with nonzero shrinkage, but the change will be two components as in Lemma 2. The effect in the ridgeless scenario is always that data mining leads to lower FPE and Vol for $\kappa''(c) < 0$ as in Corollary 1. The explicit shrinkage effect doesn't admit a closed form characterization, but is mainly driven by changes in shrinkage. When the effect from shrinkage is small comparing to the change in the ridgeless case, for example as shown in Figure 10, the behavior of FPE and Vol is similar to the ridgeless case, and from Corollary 1 we have

$$\frac{dVol}{dP_1} < 0\ ; \quad \frac{dFPE}{dP_1} < 0 \tag{65}$$

as shown in Figure 6.

B Proofs

In this section we provide proofs to theoretical propositions in the main text.

Proof of Proposition 3.1 The proposition follows directly from Lindley and Smith (1972). □

Proof of Lemma 1 When $P_1/T \to 0$, we have $\frac{1}{T}X'_{1,T}X_{1,T} \to \sigma_x^2\mathbf{I}_{qP_1}$, $\frac{1}{T}W'_TW_T \to \sigma_x^2\mathbf{I}_{(1-q)P_1}$, and $\frac{1}{T}X'_{1,T}W_{1,T} \to 0$. Therefore,

$$\hat{\beta} \to \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} + \begin{bmatrix} \frac{1}{T}X'_T\epsilon_{T+1} \\ \frac{1}{T}W'_T\epsilon_{T+1} \end{bmatrix} \to \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} \tag{66}$$

We also have

$$\frac{1}{T}\sum_{\tau}(\beta'_2X_{2,\tau})^2 \to E[(\beta'_2X_2)^2] = tr(E[\beta_2\beta'_2X_2X'_2]) = (1-q)\sigma_x^2b_*$$

The posterior variance thus becomes

$$\hat{V}_\beta \to \frac{(\sigma^2 + b_*\sigma_x^2(1-q))}{\sigma_x^2T} \to 0 \tag{67}$$

Therefore, $E[f_{t+1}|\mathcal{I}_I] = X_{1,0}\beta_1$ and $V[f_{t+1}|\mathcal{I}_I] = \sigma^2 + b_*\sigma_x^2(1-q)$. The rest of the results follow from direct calculation. □

Proof of Proposition A.1 Denote $\hat{\Psi}_T \equiv \frac{1}{T}X'_TX_T$, we can write

$$V_I[\theta] = \sigma^2X'_0(\tau I + \hat{\Psi}_T)^{-1}\frac{1}{T}\hat{\Psi}_T(\tau I + \hat{\Psi}_T)^{-1}X_0 \tag{68}$$

From Lemma 4 and the fact that $X_0 \sim N(0, \sigma_x^2\mathbf{I}_P)$, we have

$$\sigma^2X'_0(\tau I + \hat{\Psi}_T)^{-1}\frac{1}{T}\hat{\Psi}_T(\tau I + \hat{\Psi}_T)^{-1}X_0 - \sigma^2\sigma_x^2\,\mathrm{tr}((\tau I + \hat{\Psi}_T)^{-1}\frac{1}{T}\hat{\Psi}_T(\tau I + \hat{\Psi}_T)^{-1}) \to 0 \tag{69}$$

in probability. At the same time, notice that

$$\begin{aligned} \mathrm{tr}((\tau I + \hat{\Psi}_T)^{-1}\frac{1}{T}\hat{\Psi}_T(\tau I + \hat{\Psi}_T)^{-1}) &= \mathrm{tr}((\tau I + \hat{\Psi}_T)^{-1}\frac{1}{T}(\tau I + \hat{\Psi}_T - \tau I)(\tau I + \hat{\Psi}_T)^{-1}) \\ \{\text{by Lemma 7}\} &= \frac{1}{T}\,\mathrm{tr}(E[(\tau I + \hat{\Psi}_T)^{-1}(\tau I + \hat{\Psi}_T)(\tau I + \hat{\Psi}_T)^{-1}]) \\ &\quad - \frac{\tau}{T}\,\mathrm{tr}(E[(\tau I + \hat{\Psi}_T)^{-1}(\tau I + \hat{\Psi}_T)^{-1}]) \\ &= \frac{1}{T}\,\mathrm{tr}(E[(\tau I + \hat{\Psi}_T)^{-1}]) - \frac{\tau}{T}\,\mathrm{tr}(E[(\tau I + \hat{\Psi}_T)^{-2}]) \end{aligned} \tag{70}$$

From Lemma 5 and notice that when $\Psi = I_P$ we have the Marchenko-Pastur distribution, we have

$$\frac{1}{T}\operatorname{tr}(E[(\tau I + \hat{\Psi}_T)^{-1}]) \to cm(\tau; c) \tag{71}$$

and hence

$$\frac{d}{d\tau}\frac{1}{T}\operatorname{tr}(E[(\tau I + \hat{\Psi}_T)^{-1}]) \to c\frac{d}{d\tau}m(\tau; c) \tag{72}$$

Notice that

$$\frac{d}{d\tau}\frac{1}{T}\operatorname{tr}(E[(\tau I + \hat{\Psi}_T)^{-1}]) = -\operatorname{tr}(E[(\tau I + \hat{\Psi}_T)^{-2}]) \tag{73}$$

Thus, we get

$$\frac{1}{T}\operatorname{tr}(E[(\tau I + \hat{\Psi}_T)^{-2}]) \to -c\frac{d}{d\tau}m(\tau; c) \tag{74}$$

Collecting terms, we have

$$V_I[\theta] \to \sigma^2\sigma_x^2(cm(\tau; c) + c\tau\frac{d}{d\tau}m(\tau; c)) \tag{75}$$

Given the mean zero assumptions in β, X and ϵ, we have $E[\beta' X_t + \epsilon_{t+1}] = E[\hat{\beta} X_t] = 0$. Therefore,

$$RP = E[f_{t+1} - p_t] = \rho Q(V[\theta|\mathcal{I}_I] + \sigma^2) \tag{76}$$

$$\to \rho Q\sigma^2(1 + \sigma_x^2(cm(\tau; c) + c\tau\frac{d}{d\tau}m(\tau; c))) \tag{77}$$

For the price volatility (variance), we have

$$Var(p_t) = Var(\hat{\beta}' X_t - \rho Q(V[\theta|\mathcal{I}_I] + \sigma^2)) \tag{78}$$

Since $V[\theta|\mathcal{I}_I]$ converges to a constant in the limit, and given $E[\hat{\beta}' X_t] = 0$, we have

$$Var(p_t) = E[(\hat{\beta}' X_t)^2] \to \operatorname{tr}(E[X_t X_t']\hat{\beta}\hat{\beta}') \tag{79}$$

$$= \sigma_x^2 \operatorname{tr}((\tau I + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T\beta + q_T)(\hat{\Psi}_T\beta + q_T)(\tau I + \hat{\Psi}_T)^{-1}) \tag{80}$$

where we define $q_T = \frac{1}{T}\sum_{t=1}^{T} X_t'\epsilon_{t+1}$. We then use the following lemma:

Lemma 3
We have $\beta'(\tau I + \hat{\Psi}_T)^{-1}q_T \to 0$ in L_2 and thus also in probability

Proof of Lemma 3 We have that $Y_T = \beta(\tau I + \hat{\Psi}_T)^{-1} q_T$ satisfies

$$\begin{aligned} E[Y_T^2] &= E[\beta'(\tau I + \hat{\Psi}_T)^{-1} q_T q_T' (\tau I + \hat{\Psi}_T)^{-1}\beta] \\ &= E[\beta'(\tau I + \hat{\Psi}_T)^{-1} \sigma^2 \frac{1}{T} \hat{\Psi}_T (\tau I + \hat{\Psi}_T)^{-1}\beta] \\ &\to b_* \sigma^2 P^{-1} \operatorname{tr} E[(\tau I + \hat{\Psi}_T)^{-1} \frac{1}{T} \hat{\Psi} (\tau I + \hat{\Psi}_T)^{-1}] \\ &\leq \tau^{-1} b_* P/(PT) \to 0 \end{aligned} \tag{81}$$

□

Given Lemma 3, we have

$$\begin{aligned} Var(p_t) &\to \sigma_x^2 \operatorname{tr}((\tau I + \hat{\Psi}_T)^{-1} (\hat{\Psi}_T \beta\beta' \hat{\Psi}_T + q_T q_T')(\tau I + \hat{\Psi}_T)^{-1}) \\ &= \sigma_x^2 \operatorname{tr}((\tau I + \hat{\Psi}_T)^{-1} (\hat{\Psi}_T \beta\beta' \hat{\Psi}_T + \sigma^2 \frac{1}{T} \hat{\Psi}_T)(\tau I + \hat{\Psi}_T)^{-1}) \\ &= \sigma_x^2 \operatorname{tr}((\tau I + \hat{\Psi}_T)^{-1} (\hat{\Psi}_T \beta\beta' \hat{\Psi}_T)(\tau I + \hat{\Psi}_T)^{-1}) + \sigma_x^2 \sigma^2 \operatorname{tr}((\tau I + \hat{\Psi}_T)^{-1} \frac{1}{T} \hat{\Psi}_T (\tau I + \hat{\Psi}_T)^{-1}) \\ &= Term1 + Term2 \end{aligned} \tag{82}$$

We have computed $Term2$ from Eqn 70, i.e.

$$Term2 \to \sigma_x^2 \sigma^2 (c m(\tau; c) + c\tau \frac{d}{d\tau} m(\tau; c)) \tag{83}$$

For $Term1$ we have

$$\begin{aligned} Term1 &= b_* \sigma_x^2 \operatorname{tr} E[(\tau I + \hat{\Psi}_T)^{-1} \hat{\Psi}_T \hat{\Psi}_T (\tau I + \hat{\Psi}_T)^{-1}] \\ &= b_* \sigma_x^2 \operatorname{tr} E[I - 2\tau(\tau I + \hat{\Psi}_T)^{-1} + \tau^2 (\tau I + \hat{\Psi}_T)^{-2}] \\ &\to b_* \sigma_x^2 (1 - 2\tau m(\tau; c) - \tau^2 \frac{d}{d\tau} m(\tau; c)) \end{aligned} \tag{84}$$

and combining $Term1$ and $Term2$ we have

$$\begin{aligned} Var(p_t) \to \mathcal{L}(\tau; c) = b_* \sigma_x^2 (1 - 2\tau(c) m(\tau(c); c) - \tau(c)^2 \frac{d}{d\tau} m(\tau(c); c)) \\ + c\sigma^2 \sigma_x^2 m(\tau(c); c) + c\tau(c) \sigma^2 \sigma_x^2 \frac{d}{d\tau} m(\tau(c); c) \end{aligned} \tag{85}$$

The out-of-sample R^2 is defined as

$$\begin{aligned} R^2 &= 1 - \frac{E[(f_{t+1} - p_t)^2]}{E[f_{t+1}^2]} \\ &= \frac{E[f_{t+1}p_t] - E[p_t^2]}{E[f_{t+1}^2]} \\ &= \frac{E[(\beta' X_t + \epsilon_{t+1})\hat{\beta}' X_t] - Var(p_t)}{E[(\beta' X_t + \epsilon_{t+1})^2]} \end{aligned} \tag{86}$$

Notice that

$$E[(\beta' X_t + \epsilon_{t+1})^2] = \operatorname{tr} E[X_t X_t']\beta\beta' + \sigma^2 = b_* \sigma_x^2 + \sigma^2 \tag{87}$$

and we calculated $Var(p_t) \to \mathcal{L}(\tau; c)$. Therefore, it remains to calculate

$$E[f_{t+1}p_t] = Cov(f_{t+1}, p_t) = E[(\beta' X_t + \epsilon_{t+1})\hat{\beta}' X_t] = E[\beta' X_t X_t' \hat{\beta}] \tag{88}$$

given the independence of ϵ_{t+1}. Now,

$$\begin{aligned} \beta' E[X_t X_t']\hat{\beta} &= \beta' E[X_t X_t'](\tau I + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta + q_T) \\ &\to \beta' E[X_t X_t'](\tau I + \hat{\Psi}_T)^{-1}(\hat{\Psi}_T \beta) \\ &\to b_* \sigma_x^2 \operatorname{tr}((\tau I + \hat{\Psi}_T)^{-1} \hat{\Psi}_T) \\ &= \to b_* \sigma_x^2 (1 - \tau m(\tau; c)) \end{aligned} \tag{89}$$

Thus, we have $Cov(f_{t+1}, p_{t+1}) = E[f_{t+1}p_t] \to \mathcal{E}(\tau; c) = b_* \sigma_x^2 (1 - \tau m(\tau; c))$, and

$$R^2 = \frac{2\mathcal{E}(\tau; c) - \mathcal{L}(\tau; c)}{b_* \sigma_x^2 + \sigma^2} \tag{90}$$

Furthermore, we have the regression coefficient

$$\delta \equiv \frac{Cov(f_{t_1}, p_t)}{Var(p_t)} \to \frac{\mathcal{E}(\tau; c)}{\mathcal{L}(\tau; c)} \tag{91}$$

Thus,

$$Var(\delta p_t) \to \frac{\mathcal{E}(\tau; c)^2}{\mathcal{L}(\tau; c)} \tag{92}$$

and we have

$$FPE = \frac{\mathcal{E}(\tau; c)^2}{\mathcal{L}(\tau; c)} \tag{93}$$

□

Proof of Corollary 1 When investor has uninformative prior and uses ridgeless regression, using Lemma 2 we have

$$\begin{aligned} Vol &= \mathcal{L}(0;c,\kappa) \\ &= \mathcal{M}(0;c,\kappa) + \mathcal{V}(0;c,\kappa) \\ &= \begin{cases} \frac{c}{1-c}(\sigma^2 + b_*\sigma_x^2(1-\kappa)) + b_*\sigma_x^2\kappa & \text{if } c \leq 1 \\ \frac{1}{c-1}(\sigma^2 + b_*\sigma_x^2(1-\kappa)) + \frac{1}{c}b_*\sigma_x^2\kappa & \text{if } c > 1 \end{cases} \end{aligned} \tag{94}$$

For large P_1 with fixed T, $c > 1$, and since $\frac{\sigma^2}{c-1}$ is decreasing in c, we have

$$\text{sign}\left(\frac{dVol}{dP_1}\right) = \text{sign}\left(\frac{dVol}{dc}\right) = \text{sign}\left(\frac{d}{dc}(\frac{1-\kappa(c)}{c-1} + \frac{\kappa(c)}{c})\right) \tag{95}$$

Since $\kappa'(c) > 0$, we have $\frac{1-\kappa(c)}{c-1}$ to be decreasing in c. When $\kappa''(c) < 0$, we have $\kappa'(c)c < \kappa(c)$, which leads to $\frac{d}{dc}\frac{\kappa(c)}{c} < 0$. Putting both terms together we have $\frac{dVol}{dP_1} < 0$.

Similarly, we have for ridgeless regression at $P_1 > T$,

$$\begin{aligned} FPE &= \frac{\mathcal{E}(0;c,\kappa)^2}{\mathcal{M}(0;c,\kappa) + \mathcal{V}(0;c,\kappa)} \\ &= \frac{b_*\sigma_x^2\frac{\kappa(c)}{c}}{(\frac{\sigma^2}{\sigma_x^2 b_*} + 1)\frac{c}{c-1}\frac{1}{\kappa(c)} + \frac{1}{c-1}} \end{aligned} \tag{96}$$

We have $\frac{dFPE}{dP_1} < 0$ if and only if

$$\begin{aligned} &\left((\frac{\sigma^2}{\sigma_x^2 b_*} + 1)\frac{c}{c-1}\frac{1}{\kappa(c)} + \frac{1}{c-1}\right)\sigma_x^2 b_x \frac{\kappa'(c)c - \kappa(c)}{c^2} \\ &-b_x\sigma_x^2\frac{\kappa(c)}{c}\left[(\frac{\sigma^2}{\sigma_x^2 b_*} + 1)\left(\frac{c\kappa'(c) - \kappa(c)}{\kappa(c)^2(c-1)} + \frac{c}{c-1}\frac{1}{\kappa(c)}\right) + \frac{1}{(c-1)^2}\right] < 0 \end{aligned} \tag{97}$$

which is equivalent to

$$\frac{\kappa'(c)c - \kappa(c)}{c^2} \leq \frac{\sigma^2}{\sigma_x^2 b_*} + 1 + \frac{\kappa(c)}{c(c-1)} \tag{98}$$

Notice that the right hand side is always positive, and when $\kappa''(c) < 0$ we have $\kappa'(c)c < \kappa(c)$, which means the left hand size is always negative. Therefore the condition is satisfied, and we conclude that

$$\text{sign}\left(\frac{dFPE}{dP_1}\right) = \text{sign}\left(\frac{dFPE}{dc}\right) < 0 \tag{99}$$

for any κ with $\kappa''(c) < 0$. □

C Auxiliary theoretical results

In this appendix, we provide some auxiliary theoretical results for dealing with large random matrices. See Kelly et al. (2021); Didisheim et al. (2023) for additional random matrix theory results in asset pricing.

Lemma 4
Suppose that $X = (X_i)_{i=1}^{P}$ with X_i independent of X_j, $E[X_i] = 0$, $E[X_i^2] = 1$, $E[X_i^4] \leq k$ for some k, and A_P is such that $||A_P||_2 = o(1)$.

$$X_t' A_P X_t = \text{tr}(A_P X_t X_t') \tag{100}$$

$$\lim_{P \to \infty} E[(X_t' A_P X_t - \text{tr}(A_P))^2] = 0 \tag{101}$$

Proof. For the equality, we notice

$$X_t' A X_t \in R \Rightarrow X_t' A X_t = \text{tr}(X_t' A X_t) = \text{tr}(A X_t X_t') \tag{102}$$

For the second equality, we define $Y_t = X_t A_P X_t$, we have

$$E[Y_t] = E[\text{tr}(A_P(X_t X_t'))] = \text{tr}(A_P E[X_t X_t']) = \text{tr}(A_P) \tag{103}$$

and hence

$$E[(Y_t - \text{tr}(A_P))^2] = Var(Y_t) = E[Y_t^2] - E[Y_t]^2 \tag{104}$$

and thus it's sufficient to prove that $E[Y_t^2] - \text{tr}(A_P)^2 \to 0$. Now,

$$Y_t = \sum_{i,j} X_i X_j A_{i,j} \tag{105}$$

and therefore

$$Y_t^2 = \sum_{i_1, j_1, i_2, j_2} X_{i_1} X_{j_1} A_{i_1,j_1} A_{i_2,j_2} X_{i_2} X_{j_2} \tag{106}$$

Therefore,

$$\begin{aligned} E[Y_t^2] &= \sum_{i_1, j_1, i_2, j_2} A_{i_1,j_1} A_{i_2,j_2} E[X_{i_1} X_{j_1} X_{i_2} X_{j_2}] \\ &= \sum_i A_{i,i}^2 E[X_i^4] + \sum_{i,j} (A_{i,j}^2 + A_{i,i} A_{j,j}) E[X_i^2 X_j^2] \\ &\leq \sum_i k A_{i,i^2} + \sum_{i,j} A_{i,j}^2 + 2 A_{i,i} A_{j,j} \\ &= (k-1) \sum_i A_{i,i^2} + \sum_{i,j}^{2} + \text{tr}(A)^2 \end{aligned} \tag{107}$$

Notice $\sum_i A_{i,i}^2 \leq \sum_{i,j} A_{i,j}^2 = ||A||_2^2$, and thus

$$|E[Y_t^2] - \mathrm{tr}(A)^2| \leq k||A||_2^2 \tag{108}$$

and the proof is complete. □

Lemma 5
Suppose $S_t = \Psi^{1/2} X_t$ *where* X_t *is independent random vectors with* $E[X_{i,t}] = E[X_{i,t}^3] = 0$ *and* $E[X_{i,t}^2] = 1$ *and it satisfies the Lindeberg condition, and* Ψ *is symmetric and positive semi-definite. Define* $\hat{\Psi}_T = \frac{1}{T}\sum_t S_t S_t'$, *we have*

$$\lim_{T\to\infty} \frac{1}{T}\mathrm{tr}((zI + \hat{\Psi})\Psi) \to \xi(z;c)$$

almost surely, where

$$\xi(z;c) = \frac{1 - zm(-z;c)}{c^{-1} - 1 + zm(-z;c)}$$

Proof. First we introduce the Sherman-Morrison formula (see Bartlett (1951)): let $\hat{\Psi}_{T,t} = \frac{1}{T}\sum_{\tau\neq t} S_\tau S_\tau'$, we have

$$(zI+\hat{\Psi}_T)^{-1} = (zI+\hat{\Psi}_{T,t})^{-1} - \frac{1}{T}(zI+(zI+\hat{\Psi}_{T,t})^{-1})^{-1} S_t S_t'(zI+\hat{\Psi}_{T,t})^{-1}\frac{1}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t} \tag{109}$$

The proof is proceed with several steps:

- Let $\hat{\Psi}_{T,t} = \frac{1}{T}\sum_{\tau\neq t} S_\tau S_\tau'$, by 109 we have

$$\begin{aligned}&(zI+\hat{\Psi}_T)^{-1}S_t = (zI+\hat{\Psi}_{T,t})^{-1}S_t\\ &-\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}\\ &=(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}\end{aligned} \tag{110}$$

- By Lemma 4,

$$P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t - P^{-1}\mathrm{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1}) \to 0 \tag{111}$$

 in probability. At the same time, by Lemma 7,

$$P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t - E[P^{-1}\mathrm{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \to 0 \tag{112}$$

almost surely. Thus,

$$P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t - E[P^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \to 0 \tag{113}$$

in probability.

- Lemma 6 implies that

$$P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}] \to m(-z;c) \tag{114}$$

Now we have

$$\begin{aligned}
1 &= P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}(zI+\hat{\Psi}_T)] \\
&= P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}]z + P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}\hat{\Psi}_T] \\
&= zm(-z;c) + P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}\frac{1}{T}\sum_t S_tS_t'] \\
&= \{\text{symmetry across } t\} = zm(-z;c) + P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_T)^{-1}S_tS_t'] \\
&= \{\text{using Sherman - Morrison 109}\} \\
&= zm(-z;c) + P^{-1}\operatorname{tr} E[(zI+\hat{\Psi}_{T,t})^{-1}S_t\frac{1}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}S_t'] \\
&= zm(-z;c) + E[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}]
\end{aligned} \tag{115}$$

Now $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \le ||\Psi||z^{-1}$ and hence is uniformly bounded. Let's pick a sub-sequence of T such that $E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1})] \to q$ for some $q>0$. By 111 we have

$$\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t} \to \frac{c^{-1}q}{1+q} \tag{116}$$

in probability and this sequence is uniformly bounded. Hence,

$$E[\frac{P^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}{1+T^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t}] \to \frac{c^{-1}q}{1+q} \tag{117}$$

and we get

$$1 - zm(-z;c) = \frac{c^{-1}q}{1+q} \tag{118}$$

Thus, the limit of $\xi(z;c) = E[T^{-1}\operatorname{tr}(\Psi(zI+\hat{\Psi}_{T,t})^{-1}]$ is independent of the sub-sequence of T and satisfies

$$1 - zm(-z;c) = \frac{c^{-1}\xi(z;c)}{1+\xi(z;c)} \tag{119}$$

and the proof is complete. □

To compute $m(-z;c)$ we can use the following

Lemma 6
Define $m_\Psi(z) = \lim_{P\to\infty}\frac{1}{P}\operatorname{tr}((\Psi - zI)^{-1})$ *and* $m(z)$ *to be the empirical counterpart* $m(z) = \lim_{P\to\infty}\frac{1}{P}\operatorname{tr}((\Psi - zI)^{-1})$*, we have*

$$m(z;c) = \frac{1}{1-c-czm(z;c)} m_\psi\left(\frac{z}{1-c-czm(z;c}\right) \tag{120}$$

Proof. See Silverstein and Bai (1995); Bai and Zhou (2008). □

Lemma 7
We have

$$P^{-1}\operatorname{tr}(Q_P(zI+\hat{\Psi}_T)^{-1}) - E[P^{-1}\operatorname{tr}(Q_P(zI+\hat{\Psi}_T)^{-1})] \to 0 \tag{121}$$

almost surely for any sequence of uniformly bounded matrix Q_P

Proof. Let E_τ denote the conditional expectation given $S_{\tau+1}, \cdots S_T$. Let also $q_T(z) = \frac{1}{P}\operatorname{tr}(zI+\hat{\Psi}_T)^{-1}Q_P$. With this notation, since $\hat{\Psi}_{T,t}$ is independent of S_t, we have

$$(E_{t-1}-E_t)[\frac{1}{P}\operatorname{tr}(zI+\hat{\Psi}_{T,t})^{-1}Q_P] = 0 \tag{122}$$

and therefore

$$\begin{aligned}
E[q_T(z)] - q_T(z) = E_0[q_T(z)] - E_T[q_T(z)] &= \sum_{t=1}^{T}(E_{t-1}[q_T(z)] - E_t[q_T(z)]) \\
&= \sum_{t=1}^{T}(E_{t-1}-E_t)[q_T(z)] \\
&= \sum_{t=1}^{T}(E_{t-1}-E_t)[q_T(z) - \frac{1}{P}\operatorname{tr}(zI+\hat{\Psi}_{T,t})^{-1}Q_P] \\
&= \frac{1}{P}\sum_{t=1}^{T}(E_{t-1}-E_t)[\operatorname{tr}(zI+\hat{\Psi}_T)^{-1}Q_P - \operatorname{tr}(zI+\hat{\Psi}_{T,t})^{-1}Q_P] \\
&= -\frac{1}{P}\sum_{t=1}^{T}(E_{t-1}-E_t)[\gamma_t]
\end{aligned} \tag{123}$$

where we defined

$$\gamma_t = \operatorname{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}Q_P\right) \tag{124}$$

Since for any symmetric positive semi-definite matrices A, B, we have $\mathrm{tr}(AB) \leq \mathrm{tr}(A)||B||$ and $\mathrm{tr}(A^{1/2}BA^{1/2}) \leq \mathrm{tr}(B)||A||$, we have

$$\begin{aligned} ||\gamma_t|| &\leq ||Q_P||\,\mathrm{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1}\right) \\ &\leq z^{-1}\,\mathrm{tr}\left(\frac{1}{T}(zI+\hat{\Psi}_{T,t})^{-1/2}S_t(I+\frac{1}{T}S_t'(zI+\hat{\Psi}_{T,t})^{-1}S_t)^{-1}S_t'(zI+\hat{\Psi}_{T,t})^{-1/2}\right) \\ &= z^{-1}\,\mathrm{tr}(B(zI+B)^{-1}) \leq z^{-1} \end{aligned} \tag{125}$$

where $B = \frac{1}{T}S_t'(zI + \hat{\Psi}_{T,t}S_t)$. Thus,

$$(E_{t-1}-E_t)[\mathrm{tr}(zI+\hat{\Psi}_T)^{-1}\Psi] = (E_{t-1}-E_t)[\gamma_t] \tag{126}$$

forms a bounded martingale difference sequence. Applying the Burkholder-Davis-Gundy inequality (Burkholder (1966)) we get

$$\begin{aligned} E[|q_T(z)-E[q_T(z)]|^\kappa] &\leq K_\kappa P^{-\kappa} E\left(\sum_{t=1}^T |(E_{t-1}-E_t)[\gamma_t]|^2\right)^{\kappa/2} \\ &\leq K_\kappa (2T/z)^\kappa P^{-\kappa/2}(\frac{P}{T})^{-\kappa/2} \end{aligned} \tag{127}$$

Almost sure convergence follows with $\kappa > 2$ from the following lemma:

Lemma 8
Suppose that

$$E[|X_T|^\kappa] \leq T^{-\alpha} \tag{128}$$

for some $\alpha > 1$ and some $\kappa > 0$, then $X_T \to 0$ almost surely.

Proof. It is known that if

$$\sum_{T=1}^\infty Prob(|X_T| > \epsilon) < \infty \tag{129}$$

for any $\epsilon > 0$, then $X_T \to 0$ almost surely. In our case, the Chebyshev inequality implies that

$$Prob(|X_T| > \epsilon) \leq \epsilon^{-\kappa} E[|X_T|^\kappa] \leq T^{-\alpha} \tag{130}$$

and the convergence follows since $\alpha > 1$. □

The proof of Lemma 7 is thus complete. □

D Additional empirical results

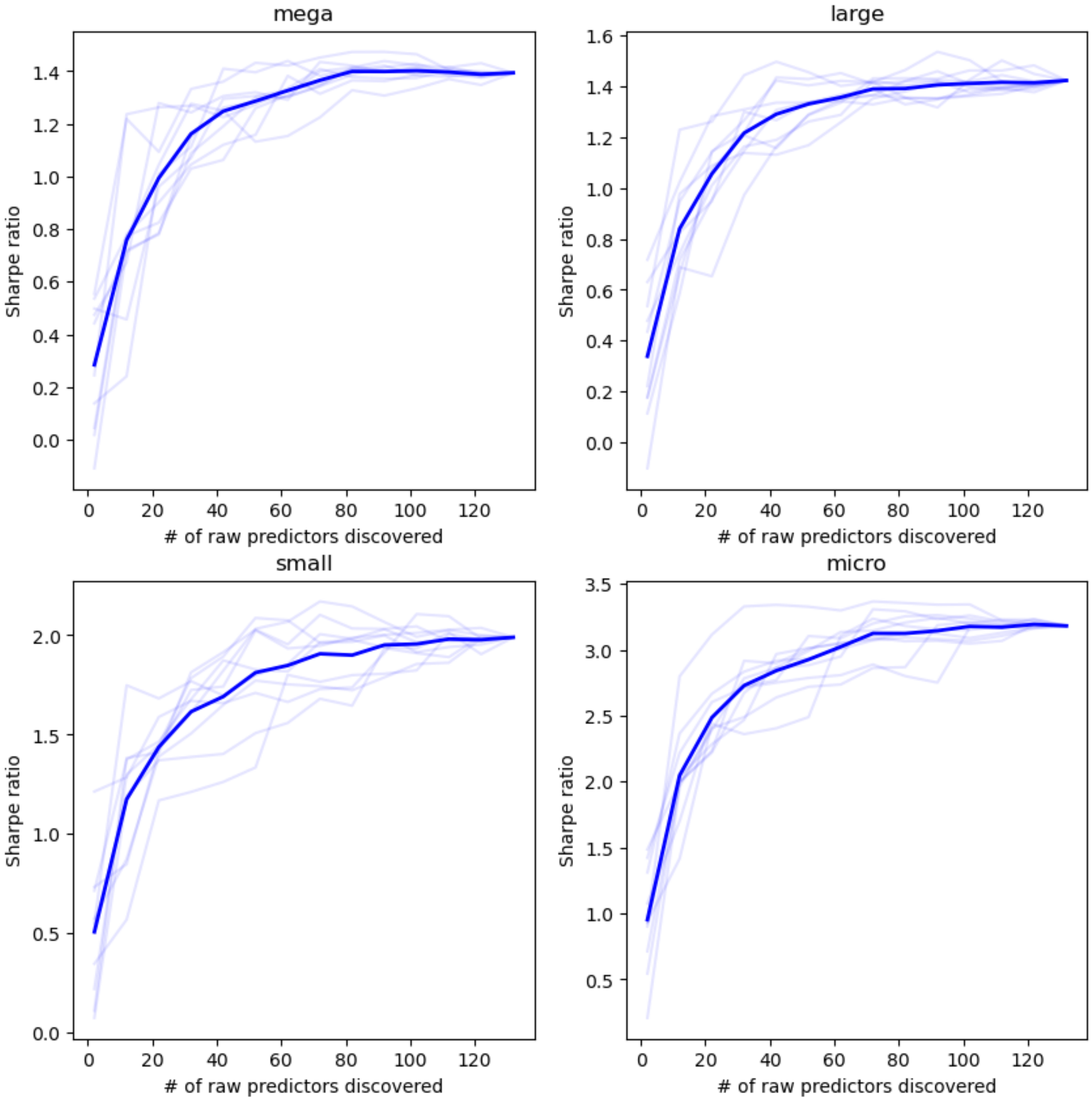


Figure 12: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed by different sizes of predictor sets. We gradually increase the set of JKP predictors used to construct factors and estimate SDF using (2) and (3). We report the highest Sharpe ratio across shrinkage z. Each light blue line represent a random order of discovering predictors, and the black line is the average across random orderings. We conduct the analysis in different market capitalization groups: mega (largest 20% of stocks based on NYSE breakpoints each period), large (between 80% and 50%), small (between 50% and 20%), and micro (between 20% and 1%).

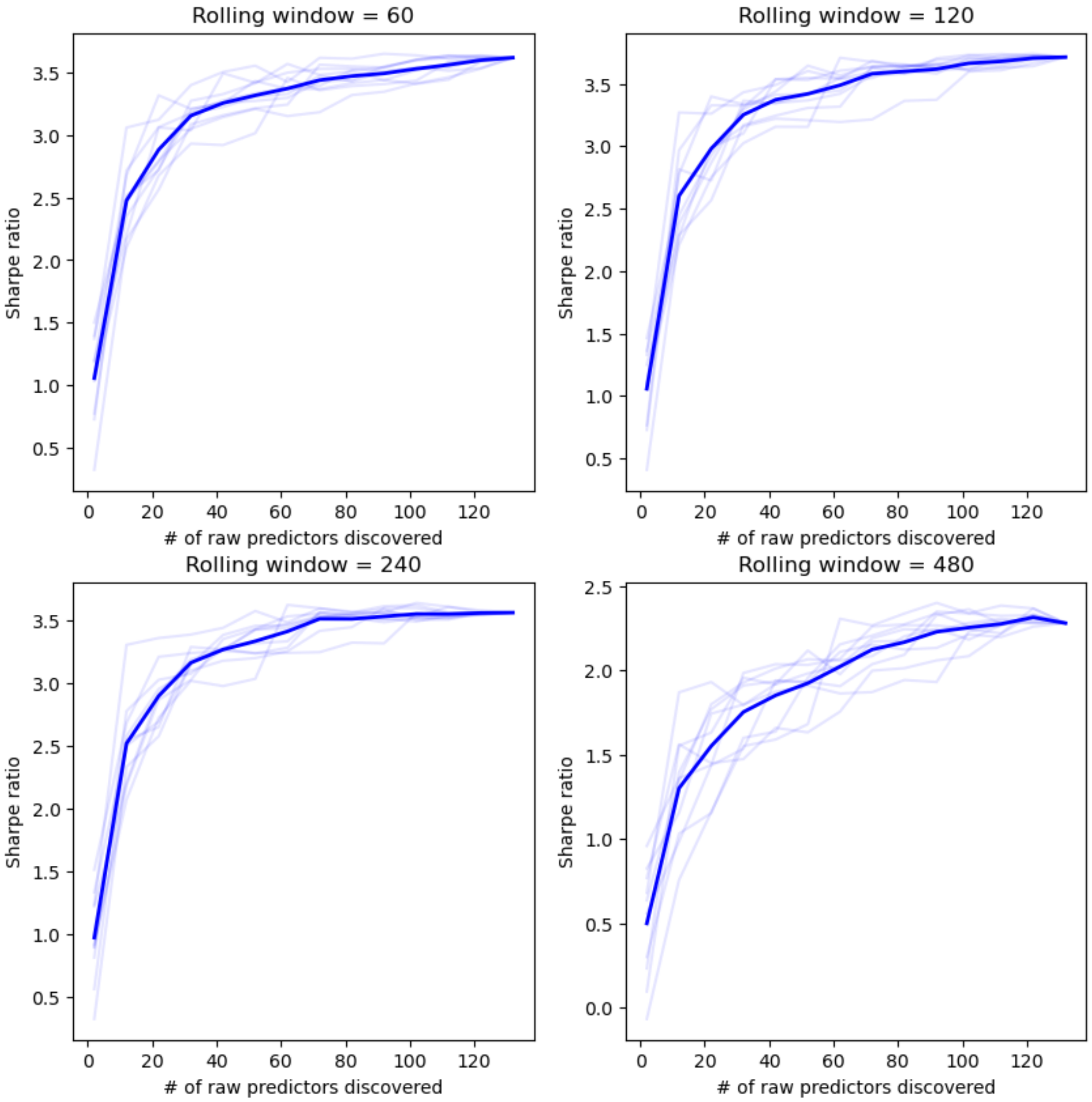


Figure 13: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed by different sizes of predictor sets. We gradually increase the set of JKP predictors used to construct factors and estimate SDF using (2) and (3). We report the highest Sharpe ratio across shrinkage z. Each light blue line represent a random order of discovering predictors, and the black line is the average across random orderings. We conduct the analysis in using different rolling window sizes (in months).

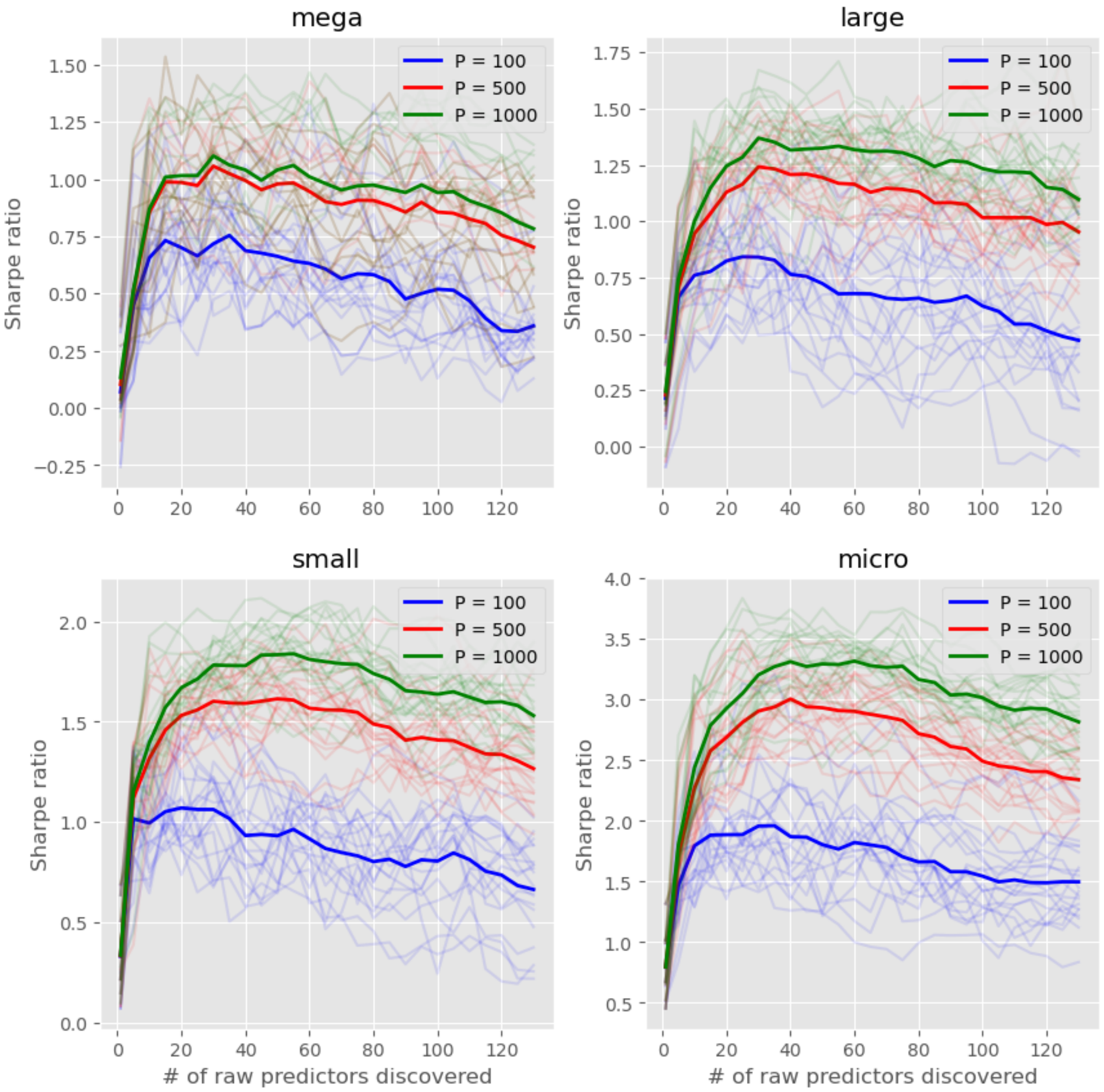


Figure 14: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed by random Fourier features (RFF) of different sizes of predictor sets. We gradually increase the set of JKP predictors used in RFF as in (4) and estimate SDF using (2) and (3) on RFFs with different size P. We average across 20 draws of random weights and report the highest Sharpe ratio across shrinkage z. Each light blue line represent a random order of discovering predictors, and the black line is the average across random orderings. We conduct the analysis in different market capitalization groups: mega (largest 20% of stocks based on NYSE breakpoints each period), large (between 80% and 50%), small (between 50% and 20%), and micro (between 20% and 1%).

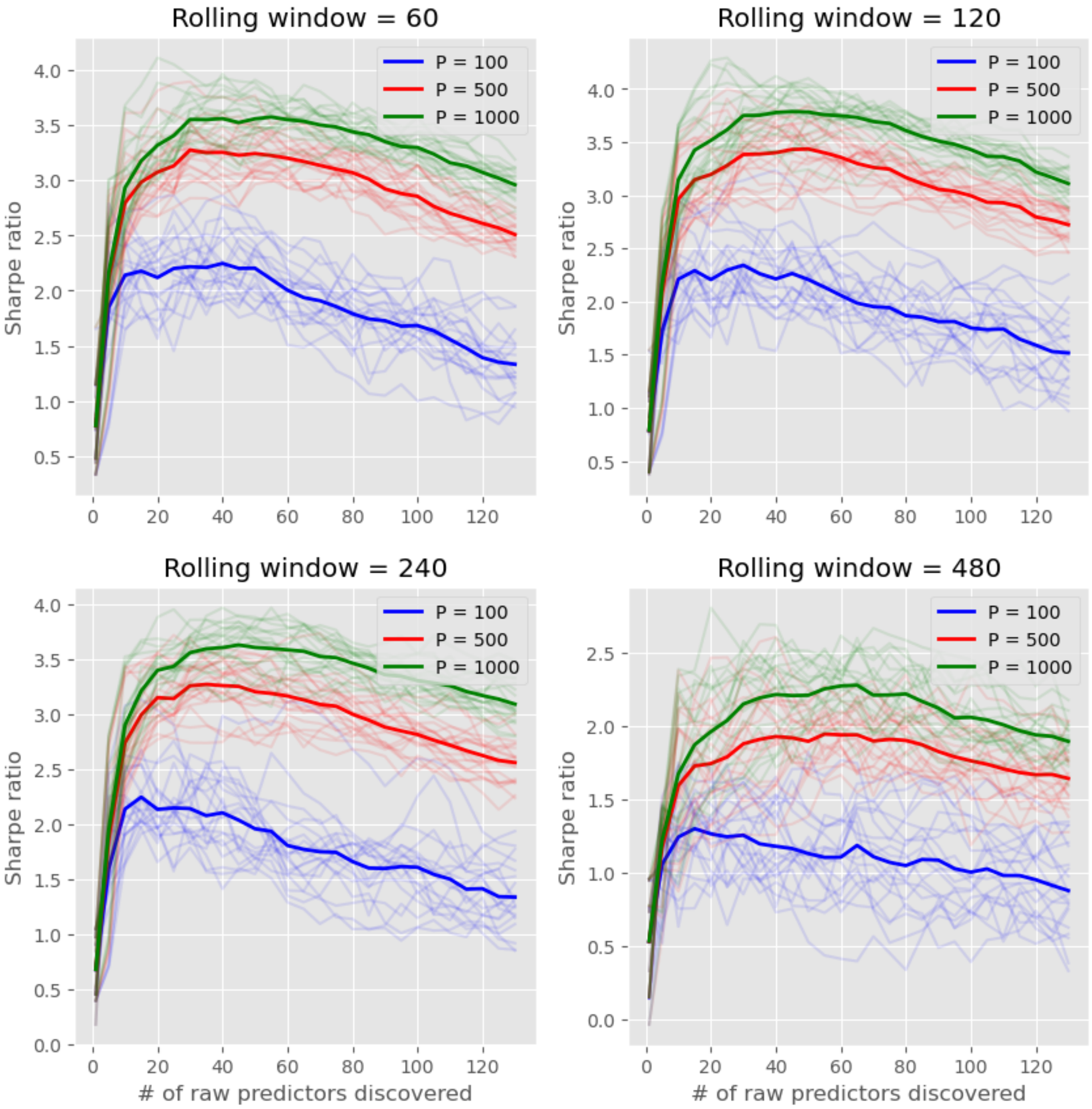


Figure 15: Out-of-sample SDF (annualized) Sharpe ratio based on factors constructed by random Fourier features (RFF) of different sizes of predictor sets. We gradually increase the set of JKP predictors used in RFF as in (4) and estimate SDF using (2) and (3) on RFFs with different size P. We average across 20 draws of random weights and report the highest Sharpe ratio across shrinkage z. Each light blue line represent a random order of discovering predictors, and the black line is the average across random orderings. We conduct the analysis in using different rolling window sizes (in months).

D.1 Discover hypothetical signals

In this section we expand the signal sets from 130 JKP signals to include more hypothetical signals, to mitigate the concern that these 130 signals are pre-selected from previous literature and are shown to have strong predictability. We construct a hypothetical signal j as the true next-period return plus noise, i.e.

$$S_{i,t}^{j} = R_{i,t+1} + \epsilon_{i,j,t}$$

with $\epsilon_{i,j,t} \sim N(0, \sigma_{\epsilon}^{2})$ and we set $\sigma_{\epsilon} = 0.05$. We conduct the signal discovery for the 130 JKP signals first, and then expand to 70 hypothetical signals. Figure 16 shows the out-of-sample Sharpe ratio of optimal portfolio constructed by expanding through 130 JKP signals and hypothetical signals. We implement the estimation in two rolling window sizes: $T = 12$ and $T = 480$. We see that even with an extremely short rolling window $T = 12$, expanding JKP signals will lead to higher performance, suggesting indeed that the information content in the JKP signals are strong. However, once we further expand to noisy hypothetical signals, we see deterioration in performance with short rolling window, while if we use a longer training window the performance will continue to increase. These results suggest that the amount of training history (and complexity) is the key driver in understanding the different behaviors when investors discover more potential predictors.

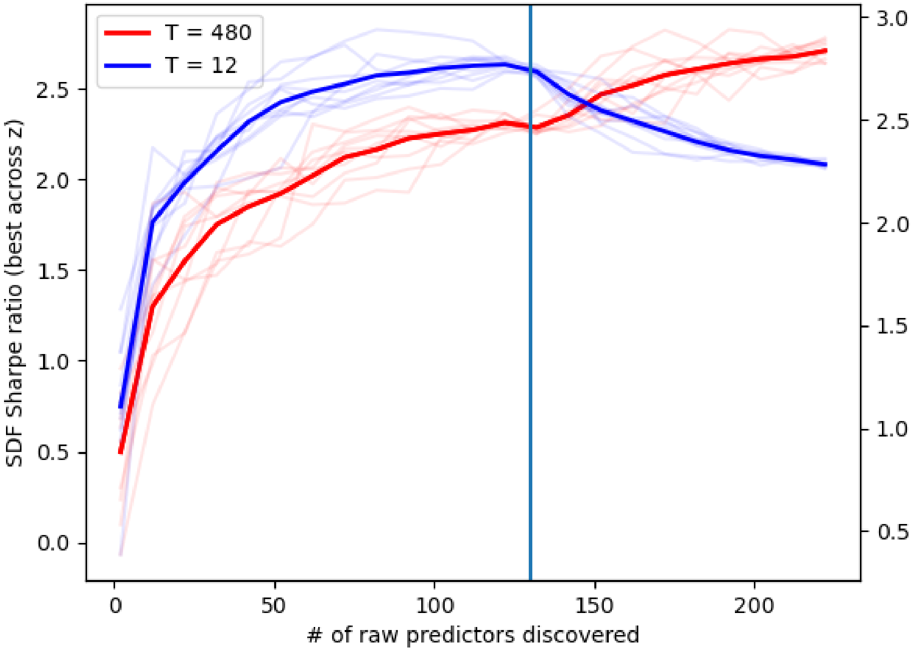


Figure 16: Hypothetical discovery of new signals. We start with the 130 JKP signals and then expend to hypothetical signals. The blue line denotes the cutoff between JKP signals and hypothetical signals.