Endogenous Misspecification*

Sebastián Guarda† José Luis Montiel Olea‡ Pietro Ortoleva§

This Version: November 11, 2025

Abstract

Decision-makers cannot consider *all* variables that may be relevant for a prediction. When should they expand their model, and when should they stop? We study agents who are aware of their potential misspecification but incur a cost when acquiring an additional variable. In a linear-regression setting, we characterize the value of adding an extra covariate. For both Bayesian and frequentist agents, we find a stronger incentive to include another covariate when their current model fits the data poorly and when few alternatives remain. Moreover, we uncover a novel form of convexity in the value of information for Bayesian agents: the more covariates already included, the greater the ex ante marginal benefit of adding another one. This result contrasts with the well-known concavity in the value of additional observations. Finally, the model implies that predictions may "jump" after unexpected evidence or theoretical developments, and behavior may appear non-Bayesian even when agents are fully Bayesian.

*We are grateful to Mira Frick, Ryota Iijima, Alessandro Lizzeri, Leeat Yariv, Sylvain Chassang, Christopher Sims, Mark Watson, Mikkel Plagborg-Møller, and participants in several seminars for their useful comments and suggestions.
†Department of Economics, Princeton University, sguarda@princeton.edu.
‡Department of Economics, Cornell University, montiel.olea@gmail.com.
§Department of Economics and SPIA, Princeton University, pietro.ortoleva@princeton.edu.

1 Introduction

Decision-makers do not always consider *all* variables that may be relevant for a prediction. Instead, they often disregard some even when they know these variables *could* matter. Central banks predicting GDP cannot include *all* potentially relevant covariates—there are simply too many. For instance, inflation expectations were included into surveys years after macroeconomic theory suggested they may play a role. Health authorities predicting outbreaks of a new disease in real time cannot quickly collect every variable that *could* matter. Likewise, an economist designing a survey or experiment must decide which variables to collect, knowing that some relevant ones will inevitably be left out. In general, complex predictions involve too many possibilities, and expanding the set of variables is costly—data must be obtained and processed, and the decision is often made *before* the new covariate can be observed, since it must first be collected.

This paper asks: How should we choose which covariates to include, and when should we expand our model? Since any time we omit a potentially relevant variable we are choosing a misspecified model, we study agents who choose *how misspecified they want to be*. Our approach complements the literature on rational inattention—by studying a specific form of inattention, namely the acquisition of covariates—and on model misspecification—by studying agents who are *aware* of their misspecification and choose it optimally in expectation. It also connects to classical results in statistics on the value of an additional data point; here, we study the value of an additional covariate. While the former is classic, the latter has received little attention.

Our model. A decision-maker must predict a variable y and wants to minimize the squared prediction error. They know that y is linearly related to a set of covariates but do not know the coefficients, as in canonical linear regressions. Suppose the agent can observe only a subset of the relevant covariates. We ask: what is the value of expanding the model and adding one more covariate, thereby reducing misspecification? We begin with a Bayesian decision-maker and later consider the frequentist counterpart.

Revising the model. Suppose first that this decision is taken after the agent has obtained n observations. We show that the value of this additional covariate is the sum of (i) its expected predictive power and (ii) the expected learning spillovers on existing variables, which *increase* with the number of covariates already included. We further show that, in large samples, the value of an additional covariate admits a simple characterization showing how the value is higher when (i) the fit of the existing model is poor and (ii) the set of remaining covariates is small.

This result has several implications. First, we should expect model "expansion" in times of poor fit—when surprising new evidence arises or when poorly fitting observations accumulate. This can

generate behavior that appears non-Bayesian even when agents are fully Bayesian: when they include a new covariate, their predictions may "jump," and an analyst unaware of the endogenous choice of variables might mistake this for a violation of Bayes' rule. Second, holding fit constant, agents will tend to enlarge their model "early on"—when few datapoints are available—and more so when their existing model is already large, since learning spillovers are then more pronounced. Third, behavior can be affected by information—including theoretical advances—that changes beliefs about *excluded* variables. For example, theoretical results showing that a currently ignored covariate is indeed irrelevant may alter behavior, since such results raise the expected explanatory power of the remaining omitted variables, and thus increase their likelihood of inclusion.

Choosing the model in advance. Next, we consider the choice of model before any data are observed—for instance, when an agent decides which variables to collect or which dataset to purchase. We show that the value of an additional covariate is not only positive—unsurprisingly for a Bayesian—but, under priors that treat covariates symmetrically, also *increasing* in the number of covariates: the larger the model, the greater the value of enlarging it further. This result holds because larger models imply learning spillovers across more variables, and this effect dominates the fact that the spillover per covariate shrinks. This "convexity" of the value of covariates contrasts with the well-known "concavity" of the value of datapoints: while the value of extra observations is positive but *decreasing*, the value of extra covariates is positive and *increasing*.

The immediate implication is that a Bayesian with symmetric priors will choose "exploding" models ex ante. Suppose each covariate has a fixed cost c, and an agent chooses how many to include. If the value of adding one covariate exceeds c, then the value of each subsequent one must also exceed c, implying that the Bayesian agent will consider the full model. Therefore, the agent either includes no covariates or all of them. Even when priors are asymmetric or costs are non-constant, our results highlight a general force pushing toward either small or very large models.

The case of the frequentist. What if the agent is a frequentist? By appropriately extending the notion of expected loss, we obtain results similar to those for Bayesians when revising a model after data: models expand when fit is poor and when few alternatives remain. However, the ex-ante results differ sharply: a frequentist may have a *negative* and *decreasing* value for additional covariates—intuitively, when the number of covariates is large relative to the number of observations. Moreover, for a class of priors in which covariates have decreasing importance, while the Bayesian may still exhibit some convexity, the frequentist never does. Finally, we study a "sophisticated" frequentist who can choose to ignore some variables if needed, and so the value of another covariate is never negative. While such an agent's values are not always positive or increasing, they are nonetheless much lower and less steep than for Bayesians.

Related Literature This paper relates to several strands of research on model misspecification, information acquisition, model selection, and non-Bayesian behavior. First of all, it contributes to the growing literature on decision-makers who are aware of their potential misspecification but choose how to manage it. Classic approaches to misspecification typically treat misspecification as exogenous—see Bohren and Hauser (2024) for a review of the literature in economic theory, and Hansen and Sargent (2008) for a review of robust control. In contrast, we study agents who are aware of omitted variables and endogenously choose how misspecified to be, trading off precision against the cost of expanding their model. Related papers that study endogenous model switching include Cho and Kasa (2015), Fudenberg and Lanzani (2023), Gagnon-Bartsch et al. (2023), He and Libgober (2025), and Ba (2025). Rather than switching to a different model as in these papers, we study agents who can reduce their misspecification at a cost, within a canonical linear regression framework—a natural setting for analyzing specification choice that yields structured predictions. Many of these papers find that the goodness of fit of the current model is a crucial determinant of whether to act, either by switching models or, in our case, enlarging one. This connects to recent experimental evidence showing that goodness of fit is an important consideration in model selection (Barron and Fries, 2024; Ambuehl and Thysen, 2024; Aina and Schneider, 2025), and helps rationalize why agents and institutions periodically enlarge their forecasting frameworks in response to poor fit, echoing patterns observed in macroeconomic forecasting (e.g., Giacomini et al., 2020) and in the evolution of econometric models in practice.

The paper also contributes to the literature on information acquisition and rational inattention, beginning with Sims (2003); see Maćkowiak et al. (2023) for a review. While rational inattention studies the optimal precision of signals under information constraints, we focus instead on the selection of covariates—a discrete, structural form of inattention. In this sense, our framework is closer to Hellwig and Veldkamp (2009) and Myatt and Wallace (2012), who analyze endogenous signal structures, but differs in modeling the model's dimension as the choice variable. It also relates to the sparse inattention formulation of Gabaix (2014, 2019) where agents may fully ignore some relevant dimension in their decisions, and Alaoui and Penta (2016) where there is a discrete choice over depth of reasoning in games.

Our analysis also connects to the statistical literature on model selection and the value of additional regressors. The decomposition of the value of a covariate into direct and spillover components parallels classical results on the bias–variance trade-off and the "value of a data point" in regression (e.g., Leamer, 1978), but our focus on covariates rather than observations yields distinct comparative statics. The finding that Bayesian agents exhibit convexity in model dimensionality complements results in Bayesian experimental design showing concavity in sample size (e.g., Lindley, 1956; Chaloner and Verdinelli, 1995). The value of another observation has also been the recent focus on a literature studying the data economy, surveyed in Veldkamp and Chung (2024).

Finally, our work speaks to a broader behavioral literature that interprets apparent "non-Bayesian" behavior as rational responses to model uncertainty. By showing how endogenous model revision can produce discontinuous updates in beliefs, our analysis offers a Bayesian microfoundation to models of non-Bayesian behavior based on model switching, like Ortoleva (2012) where, like in our model, agents reconsider their approach if their current one fits the observed data poorly, selecting a new model from a prior over priors.

2 The Environment

A decision-maker needs to make a forecast $\hat{y}_n$ of a variable of interest $y_n \in \mathbb{R}$ to minimize the square forecast error $(y_n - \hat{y}_n)^2$. The decision-maker knows that each y_i is drawn independently from the following Data Generating Process (DGP):

$$y_i = x_i\beta \qquad x_i \sim \mathcal{N}_k\left(0, I_k\right),$$

where I_k is the identity matrix of size k. For each i, x_i is drawn i.i.d.. This structure is known to the agent, except that $\beta \in \mathbb{R}^k$ are (typically) unknown.

Three assumptions are implicit in this DGP. First, it assumes independent covariates, for simplicity; our results easily extend to the more general case of correlation, as we discuss in Section 5. Second, it assumes a linear structure, but this is not very restrictive (once correlation is allowed): since all functions of a variable are permitted as covariates themselves (e.g., the square or cube of a covariate), many functional forms can be approximated arbitrarily well. Third, it posits that the complete set of covariates captures *all* the variation y—there is no additional "pure noise." While our results immediately generalize to adding extra noise, this formulation conceptually captures the idea that a complete model would leave no space for uncertainty. Naturally, however, variation akin to noise will emerge once incomplete models are considered, as we will see momentarily.

Models and Data. To help with their forecast, the agent has access to a set of covariates and a dataset of n past observations. The critical assumption is that they have access only to a *subset* $J \subset \{1, ..., k\}$ of the covariates. Denoting x_{Jt} as the corresponding vector for each $t \in \{0, \ldots, n\}$ (and x_{-Jt} the covariates left out), to predict y_n agents have access to x_{Jn} as well as a dataset with n observations: $D_J := \left\{\{y_i, x_{Ji}\}_{i=0}^{n-1}\right\}$. If we let β_J denote the corresponding regression coefficients and β_{-J} those of the covariates not included, it is immediate to see that the agent will conduct their forecasts as if the DGP were

$$y_i = x_{Ji}\beta_J + \varepsilon_{Ji} \qquad \varepsilon_{Ji} \sim \mathcal{N}\left(0, \sigma_J^2\right)$$

where $\varepsilon_{Ji} = x_{-Ji}\beta_{-J}$ and $\sigma_J^2 = \sum_{j \notin J} \beta_j^2$. (Note that the latter holds because covariates are assumed to be independent.) Intuitively, the agent does not observe the covariates outside of J, and treats them as part of a noise term; indeed, the usual noise term could be understood as the outcome of unobserved covariates.

In what follows, we use the expression "model J" to denote the family of plausible joint distributions of y_i and x_{Ji} coming from the subset J. That is, given $\theta_J := \left\{\beta_J, \sigma_J^2\right\}$, model J is given by $\left\{Q_{\theta_J}\right\}_{\theta_J \in \mathbb{R}^{|J|} \times \mathbb{R}_+}$ where Q_{θ_J} denotes the joint distribution of y_i and x_{Ji} induced by parameters θ_J.

The Choice of Models. The bigger J, the more information in D_J—adding another column to the dataset —, and, therefore, the better the predictions in expectation. The goal of this paper is to quantify *how much* better they are: What are the incentives to expand the set J?

As discussed in the introduction, we consider two possible timings. In the "ex-post problem," the agent observes a dataset with n observations and J covariates first, and then decides whether to expand the model to obtain data from a larger dataset $J' \supseteq J$. In the "ex-ante problem," we study the decision of which J covariates to observe, knowing that n observations will be available and assuming no data can be observed later.

In both cases, our task is to characterize the benefit of adding an additional covariate to the model. Then, the agent will need to compare it to the cost of including the covariate, for a given cost function $c(J, n)$. Together, this gives all we need to characterize the optimal J.

Priors and Large Samples. For Bayesian decision-makers, let π denote the prior over the unknown parameters of the DGP, updated using D_J. As it is well known, analytical results for posterior distributions are notoriously hard to obtain unless one assumes a conjugate prior. Unfortunately for our setup, if a prior is conjugate for one model J, then generically it will not be conjugate for another J'.[1] To obtain analytical results, we will at times focus on large samples, and to this end, posit a high-level assumption on the prior π that guarantees that the posterior variance decreases at rate n.

Assumption 1. *As $n \to \infty$, for any $j \in J$, we have $n\mathbb{V}_\pi\left[\beta_j \mid D_J\right] \xrightarrow{p} \sigma_J^2$.*

This property holds under mild regularity conditions, as shown by the Bernstein-von Mises

[1]For model J to have a conjugate prior, σ_J^2 must follow an inverse gamma distribution. But if we also want a conjugate prior for model $J' = J \cup \{l\}$, then $\sigma_{J'}^2 = \sigma_J^2 + \beta_l^2$ must also follow an inverse gamma. We are not aware of any distribution for β_l^2 that would result in $\sigma_{J'}^2$ following an inverse gamma (outside of a degenerate distribution at 0).

theorem.[2] If prior variance is finite and atomless, Assumption 1 implies that, as $n \to \infty$,[3]

$$\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right] = \mathbb{E}_\pi\left[\sigma_J^2\right]/n + o\left(n^{-1}\right). \tag{1}$$

We refer to the "asymptotic approximation" as the value of $\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right]$ (and functions of it) that disregards the remainder term $o\left(n^{-1}\right)$. Appendix B.1 demonstrates that it is quite accurate, even for relatively small samples.

3 Endogenous Misspecification for Bayesian Decision-Makers

3.1 Revisiting the Model After Data

We begin our analysis with the case in which agents have observed a dataset with covariates J and need to decide whether to enlarge it with a new covariate $\{x_{li}\}_{i=0}^n$ at a given cost. If they do, they enrich their dataset with this new covariate for all observations. Let D_J denote the original dataset, $J' = J \cup \{l\}$ the larger set of covariates, and $D_{J'} = \left\{\{y_i, x_{J'i}\}_{i=0}^{n-1}\right\}$ the larger dataset.

After observing dataset D_J, a Bayesian agent updates their prior and forms their forecast of y_n as a function of the observed covariates x_{Jn}. The optimal prediction is the one that minimizes the expected loss; denote it by $\hat{y}_B(x_{Jn})$. A standard and intuitive result in Bayesian statistics is that $\hat{y}_B\left(x_{Jn}\right) = x_{Jn}\mathbb{E}_\pi\left[\beta_J \mid D_J\right]$.

For our purposes, it is useful to characterize the loss that the agent expects to have after observing D_J but before observing x_{Jn}, that is

$$L_B\left(D_J\right) := \mathbb{E}_\pi\left[\mathbb{E}_{Q_\theta}\left[\left(y_n - \hat{y}_B\left(x_{Jn}\right)\right)^2\right] \mid D_J\right].$$

We refer to this object as the *expected posterior loss*. The following lemma characterizes it.

[2]The Bernstein-von Mises theorem shows that the posterior $\beta_J \mid D_J$ satisfies

$$\left\|\beta_J \mid D_J - \mathcal{N}_{|J|}\left(\hat{\beta}_J, \frac{\sigma_J^2}{n} I_{|J|}\right)\right\|_{TV} \overset{Q_{\theta_J}}{\to} 0,$$

where $\|\cdot\|_{TV}$ denotes the total variation distance and $\hat{\beta}_J$ is the OLS estimator of β_J, establishing an asymptotic relationship between Bayesian and frequentist statistics: the Bayesian credible interval and the frequentist confidence interval coincide in large samples, which means the Bayesian credible interval shrinks at rate n, implying our assumption.

[3]Eq. (1) can be shown as follows. First, under Q_π, the sequence of expected posterior variances is strictly decreasing and therefore bounded, as can be shown with the same arguments as in the proof of Proposition 2. Second, if the prior variance is atomless, the posterior variances are atomless as well. Therefore, this sequence is uniformly integrable. Thus, by the Vitali convergence theorem, it converges in mean, and so $n\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right] \to \mathbb{E}_\pi\left[\sigma_J^2\right]$, implying (1).

Lemma 1. *The following is true:*

$$L_B(D_J) = \sum_{j \in J} \mathbb{V}_\pi \left[\beta_j \mid D_J\right] + \sum_{j \notin J} \mathbb{E}_\pi \left[\beta_j^2 \mid D_J\right]. \tag{2}$$

Intuitively, the expected posterior loss can be decomposed into two terms. First, $\sum_{j \in J} \mathbb{V}_\pi \left[\beta_j \mid D_J\right]$: the uncertainty coming from imperfect knowledge of β_J, the coefficients for the *observed* covariates; this is what Montiel Olea et al. (2022) denote "model estimation uncertainty". Second, $\sum_{j \notin J} \mathbb{E}_\pi \left[\beta_j^2 \mid D_J\right]$: the uncertainty coming from the fact that some covariates are *unobserved*.

With this characterization, we can turn to the value of including a new covariate: this will be the expected decrease in expected posterior loss, which must be evaluated before the new covariate is observed. The value of an additional covariate l is therefore

$$\mathcal{V}_B(D_J, l) := L_B(D_J) - \mathbb{E}_{Q_x}\left[L_B(D_{J'}) \mid D_J\right], \tag{3}$$

where the expectation is over $D_{J'} \backslash D_J = \{x_{li}\}_{i=0}^{n-1}$, with Q_x denoting its distribution. Our first result shows that also $\mathcal{V}_B(D_J, l)$ admits a simple characterization with an intuitive interpretation.

Lemma 2. *The following is true:*

$$\mathcal{V}_B(D_J, l) = \mathbb{E}_\pi\left[\beta_l \mid D_J\right]^2 + \sum_{j \in J'} \left(\mathbb{V}_\pi \left[\beta_j \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi \left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right). \tag{4}$$

The expected gains from enlarging the dataset are the sum of two terms, which we can understand as a direct and an indirect effect. The first term, $\mathbb{E}_\pi\left[\beta_l \mid D_J\right]^2$, captures the direct effect: the additional covariate is directly useful for making predictions, and this is particularly true the more we expect it to matter—if $\beta_l = 0$, it won't help, if $|\beta_l|$ is large, it will. The second term, $\sum_{j \in J'} \left(\mathbb{V}_\pi \left[\beta_j \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi \left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right)$, captures the indirect effect: how much is observing one additional covariate improving my ability to understand the *existing* ones, a spill-over learning effect. Another way to understand these effects is to think of including a covariate as moving some of the variation from the unobserved noise term ε_J into an observed element, thereby reducing the noise in the estimation σ_J^2. A direct effect is that this reduces unexplained variation, while the indirect effect is that it also *speeds up learning* about the model's parameter estimates for the covariates that are already observed.

Relation to Model Fit in Large Samples. How does the benefit of an additional covariate relate to how the previous model performed? To study this case, we focus on large samples and on Dirichlet priors that include only parameters compatible with the observed noise. In particular, recall that by Assumption 1, the decision-maker eventually learns the regression coefficients inside the model

and precisely estimates σ_J^2, but receives no information about which of the covariates outside the model contributes to it. Then, let $\tilde{\Theta}_j$ denote all distributions of normalized coefficients β_j for $j \notin J$ compatible with σ_J^2, that is:

$$\tilde{\beta}_j := \frac{\beta_j}{\sigma_J} \quad \forall j \notin J \qquad \tilde{\Theta}_J := \left\{ \tilde{\beta}_j^2 : \sum_{j \notin J} \tilde{\beta}_j^2 = 1 \right\}.$$

Proposition 1. *If π is such that $\left\{\beta_j^2\right\}_{j\notin J}$ jointly follow a Dirichlet distribution with parameter vector $\left\{\alpha_j\right\}_{j\notin J}$ with support in $\tilde{\Theta}_J$, then as $n \to \infty$*

$$\mathcal{V}_B\left(D_J, l\right) \xrightarrow{p} \frac{\alpha_l}{\sum_{j\notin J} \alpha_j} \sigma_J^2.$$

Intuitively, the value of an additional covariate in large samples is the share of the residual noise that can be attributed to that variable according to the agent's *prior*. When that is symmetric for all varibles left out ($\alpha_j = \alpha_i$ for all $i, j \notin J$), the value of each additional covariate converges to $\frac{\sigma_J^2}{k-|J|}$.[4] Interestingly, this can be directly related to how well model J fits the data. Specifically, consider the population coefficient of determination, which captures how much of y's variance is explained by model J

$$R_\infty^2\left(J\right) := 1 - \frac{\sigma_J^2}{\mathbb{V}\left[y_i\right]}.$$

Corollary 1. *Under the conditions of Proposition 1*

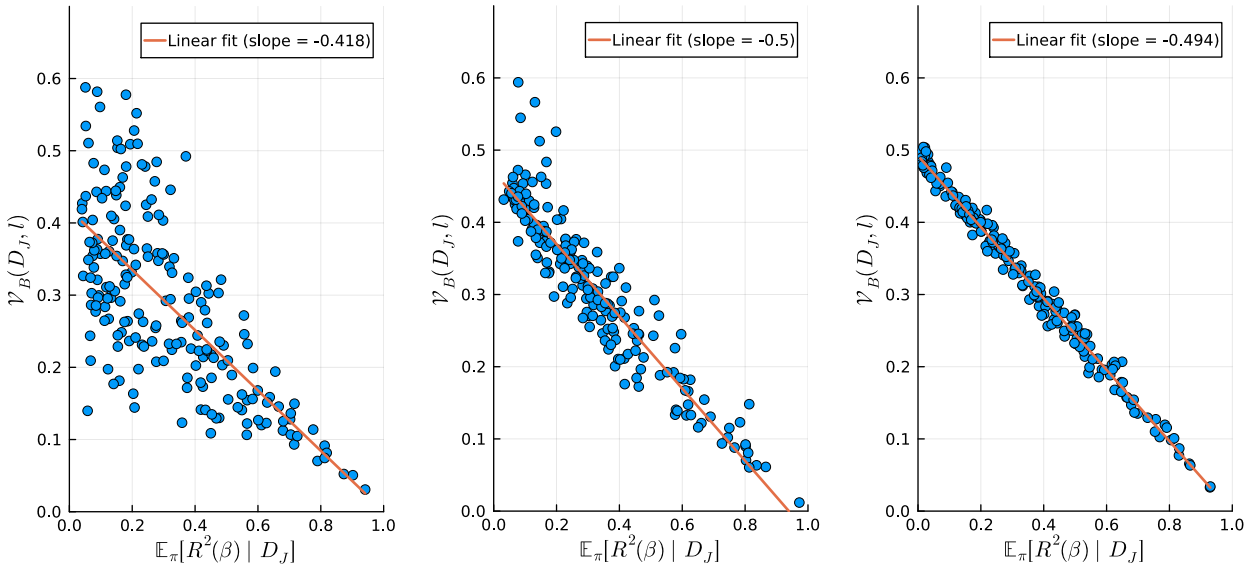
$$\mathcal{V}_B\left(D_J, l\right) \xrightarrow{p} \frac{\alpha_l}{\sum_{j\notin J} \alpha_j} \left(1 - R_\infty^2\left(J\right)\right) \mathbb{V}\left[y_i\right]$$

This result shows that the value of an additional covariate has a simple formula. Holding fixed the variance of y, which scales the value of learning, the value of an additional covariate is i) decreasing in model fit and ii) increasing in the expected share of explanatory power. Both are intuitive. If model fit is high, the variation left to be explained is small, and so is the benefit of an additional covariate. And if the agent believes most of the observed noise can be attributed to other covariates, the expected benefit is small. This is even more explicit in the symmetric case, where the formula becomes

$$\mathcal{V}_B\left(D_J, l\right) \xrightarrow{p} \frac{1}{k - |J|} \left(1 - R_\infty^2\left(J\right)\right) \mathbb{V}\left[y_i\right].$$

When many covariates are left out and the agent is unsure which are relevant, the expected value of an additional covariate decreases, as the agent is worried it may not be relevant.

[4] It is easy to see how this latter result does not rely on the Dirichlet functional form—all we need is a symmetric prior over the β_js for $j \notin J$.



(a) $n = 10$, all simulations (b) $n = 10$, only $\hat{\sigma}_y^2 \approx 1$ (c) $n = 100$, only $\hat{\sigma}_y^2 \approx 1$

Figure 1: Model Fit and Value of an Additional Covariate

Small Samples. While solving our problem analytically for small samples is challenging since we do not have conjugate priors, simulations suggest that the essence of our result holds in that case as well. In particular, we continue to find a negative relation between the value of an additional covariate and model fit. We discuss our main results here and leave details to Appendix B.3.[5]

Figure 1 plots the value of an additional covariate and our notion of model fit for many different values of β and $\{x_i\}_{i=0}^n$. Panel (a) shows a clear negative relationship when $n = 10$, where each dot represents a different β and $\{x_i\}_{i=0}^n$. Part of the variation in this graph stems from the heterogeneous empirical variances of y across small-sample simulations—if y appears volatile, a Bayesian agent infers a higher value for a new covariate. Panel (b) shows that the relationship is much tighter if we focus on simulations in which the estimated variance lies within the interval [0.999,1.001]. Increasing observations to $n = 100$ makes this even more so, as shown in Panel (c).

3.1.1 Implications

Change The Model When Forecasts Fail. Our results immediately imply we should observe "expansions" of the model when forecasting is poor, e.g., with a string of observations incompatible with the current model. This may happen, for example, when surprising new evidence comes to

[5]To define expected model fit for a Bayesian in small samples, we follow Gelman et al. (2019). Denote by R^2 the ratio explained variance over total variance, $R^2(\beta) := \frac{\mathbb{V}_{Q_\beta}[x_{Ji}\beta_J]}{\mathbb{V}_{Q_\beta}[y_i]} = \frac{\sum_{j\in J}\beta_j^2}{\sum_{j=1}^k \beta_j^2}$, where Q_β is the joint distribution of $\{x_i, y_i\}$ implied by β, and the equality follows immediately from the definition of y_i and the distributions of x_i. We can then define the expected coefficient of determination as $\mathbb{E}_\pi\left[R^2(\beta) \mid D_J\right]$.

light—reminiscent of several epistemological models that explain how new evidence may spur scientific "revolutions." In our setup, this may occur even at a much smaller scale: a sequence of observations that is compatible with the existing model, yet fits it poorly, may prompt the decision-maker to consider a new covariate—one that could induce radically different predictions.

In turn, this means we may observe discrete "jumps" in forecasting, because the agent will have an additional $x_{ln}\mathbb{E}_\pi\left[\beta_l \mid D_J\right]$ term in the forecast $\hat{y}_n$. This may happen after an "extreme" observation that prompts the need for a new model, or even an observation not far from the model's current predictions, but that pushes the model fit just below the threshold—the "last straw"—inducing the need for a new covariate and, possibly, a substantial change in predictions.

This also suggests a lower bound in model fit across problems and across agents. Recall that our results show that (for large enough samples), $\mathcal{V}_B\left(D_J, l\right) \approx \frac{1}{k-|J|}\left(1 - R^2_\infty\left(J\right)\right)\mathbb{V}\left[y_i\right]$. Assume that the cost of acquiring a new covariate is the same for each one left out, and call it $\bar{c}$. As the agent will acquire a new covariate if $\mathcal{V}_B\left(D_J, l\right) \geq \bar{c}$ for some $l \notin J$, we must have $\mathcal{V}_B\left(D_J, l\right) < \bar{c}$ for the model actually used in the long run, giving us a lower bound on model fit:

$$R^2_\infty\left(J\right) > 1 - \bar{c}\frac{k - |J|}{\mathbb{V}\left[y_i\right]}.$$

Finally, an economic implication is that agents may use the type of information acquisition we study to dampen the effects of shocks that increase uncertainty, with important consequences for predicting their responses. For example, if we observe increased volatility in firm-level outcomes (e.g., revenue), we may be wary of the negative effect on investment due to increased uncertainty, as in Bloom (2009). However, realized volatility is not exactly the same as perceived uncertainty, and in this relationship, it's important to account for potential endogenous information acquisition. If increased volatility leads to model expansion, as our theory predicts, then the effect on perceived uncertainty and investment might be reduced. In fact, if—by chance—the newly included covariate proves to be particularly effective in explaining variation, to the point that the new model's fit is higher than the previous model's fit before the increase in volatility, then we may even observe *reverse* effects—investments may increase despite the volatility increase.

Change The Model Early On and When the Model is Large. Lemma 2 shows that the benefit of a new covariate is the sum of the variance explained by that new covariate and the spill-over learning effects on other covariates. This means that the acquisition of a new covariate will tend to happen particularly when, fixing model fit, little data is available and spill-over learning effects are more valuable (as the βs for existing covariates are estimated with more noise, that is, $\mathbb{V}_\pi\left[\beta_j \mid D_J\right]$ is high).

This also means that model change will tend to occur when, holding model fit constant, the

existing model already has many covariates, leading to greater spill-over learning. That is, fixing fit, model expansion tends to occur early on, and when models are already larger to begin with.

Change The Model When You Know What to Add: The Role of Theory. An additional implication of our results is intuitive: agent expand their model only when they know what to add. When too many alternatives are available and agents are unsure of which are the important ones, they will be wary of incurring the cost, given the risk that the new covariate will prove useless.

While this is intuitive, it implies that new information about the variables *excluded* from the model may also affect behavior. First, information in support of an excluded covariate will, naturally, matter. For example, new theoretical work may suggest that a previously disregarded covariate could have strong predictive power. Our results show that if this change is large enough, the agent will actually include this new covariate, following a simple formula. A bit less directly, also information fully compatible with the current model may induce change. Suppose, for example, the agent is currently disregarding covariates x_7, x_8, and x_9, and that their prior about their role is symmetric. Suppose that new information arises, like a new theoretical model, that shows that x_9 is indeed irrelevant ($\beta_9 = 0$), confirming the agent's decision to leave it out. Yet, this may also induce a discrete jump in behavior: as the agent now has better information about the role of the *other* variables left out, they may choose to include at least one of them, since each is (in expectation) now explaining half of the residual variance instead of only one third.

A Bayesian Explanation for Non-Bayesian Behavior? The implications above suggest that our framework may provide a Bayesian explanation for what may appear to be non-Bayesian behavior. If the analyst does not observe the agent's model and mistakenly assumes either that all covariates are used or that the model is unchanged, they may believe that the Bayesian agent we study is, in fact, non-Bayesian. This is clear in the examples above: major shifts in predictions due to very noisy additional information compatible with the existing models—like the "last straw" example discussed above—can easily be mistaken for violations of Bayes' rule. Indeed, this behavior is reminiscent of the model in Ortoleva (2012), where the agent's "reconsiders" the model they are using if the model fit is sufficiently low.

This suggests a decomposition of forecast updates. Consider a Bayesian agent with model J that observes x_{Jn} before observing D_J. The forecast in this case will use the prior belief over β_J:

$$\hat{y}_n = x_{Jn}\mathbb{E}_\pi\left[\beta_J\right]$$

Consider the case where, after they observe D_J, they expand the model to $J \cup \{l\}$, updating the forecast to

$$\hat{y}'_n = x_{J'n}\mathbb{E}_\pi\left[\beta_{J'} \mid D_{J'}\right]$$

Then, the forecast update can be decomposed as follows:

$$\hat{y}'_n - \hat{y}_n = \underbrace{x_{Jn}\left(\mathbb{E}_\pi\left[\beta_J \mid D_J\right] - \mathbb{E}_\pi\left[\beta_J\right]\right)}_{\text{"Bayesian" update}} + \underbrace{x_{Jn}\left(\mathbb{E}_\pi\left[\beta_J \mid D_{J'}\right] - \mathbb{E}_\pi\left[\beta_J \mid D_J\right]\right) + x_{ln}\mathbb{E}_\pi\left[\beta_{J'\setminus J} \mid D_{J'}\right]}_{\text{"Non-Bayesian" update}}.$$

The first term in the non-Bayesian update captures the improved estimation of β_J, and the second term references the inclusion of a new predictive covariate. Despite this apparent deviation from standard Bayesian norms, of course, this behavior has a fully Bayesian rational underpinning.

Empirical Evidence. Several papers across fields documented an increase in information acquisition in times when agents' forecasting accuracy decreases. For investors deciding whether to acquire stocks, Andrei et al. (2023) find that heightened macro and micro economic uncertainty increases downloads at SEC EDGAR, a public repository for company SEC filings, and company-specific google searches; Fischer et al. (2024) find that EDGAR downloads and Bloomberg terminal attention for a firm are higher when that firm's observed price deviates more from the price that would be expected based on earnings. In a study of professional forecasters during the 2008 financial crisis, Giacomini et al. (2020) shows behavior reminiscent of model revision in light of poor model fit; narrative evidence suggests similar behavior by governments and international institutions during the COVID pandemic (Buell et al., 2021; Barbaglia et al., 2023).

3.2 Choosing the Model Ex-Ante

What if the agent needs to decide which covariates to observe *before* seeing any data? This is the case, for example, of a researcher, company, or institution that needs to choose which dataset to purchase or assemble before accessing any of it, or provide company-wide guidelines on how to approach each problem. Specifically, suppose that the decision-maker chooses the model J before observing any data subject to a cost that is increasing model size, receives dataset $D_J = \left\{\{y_i, x_{Ji}\}_{i=0}^{n-1}\right\}$, observes x_{Jn}, and, finally, makes a prediction about y_n. How is J chosen?

As in the previous section, the Bayesian agent has a prior over the unknown parameters of the DGP, which is updated with the data and used to form a loss-minimizing optimal prediction. While the characterization of the loss conditional on D_J in Lemma 1 continues to hold, here we need to compute the expectation before D_J is realized. Denoting by Q_π the joint distribution over D_J implied by π, this is given by

$$\mathcal{L}_B\left(J, n\right) := \mathbb{E}_{Q_\pi}\left[L_B\left(D_J\right)\right] = \sum_{j \in J} \mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right] + \sum_{j \notin J} \mathbb{E}_\pi\left[\beta_j^2\right]. \tag{5}$$

If $c\left(J, n\right)$ represents the cost of acquiring dataset D_J of length n, the model/data decision is

given by a standard tradeoff:

$$\min_{J,n} \mathcal{L}_B(J,n) + c(J,n).$$

Our task here will be to characterize $\mathcal{L}_B(J,n)$, the new object of this analysis.

Monotonicity of $\mathcal{L}_B$. A first, trivial observation is that $\mathcal{L}_B$ must be decreasing in both arguments—both bigger n and a larger set of covariates mean more data and must entail a weakly lower expected loss, that is, $J' \supseteq J$ and $n' \geq n$ imply $\mathcal{L}_B(J,n) \geq \mathcal{L}_B(J',n')$. To characterize how $\mathcal{L}_B(J,n)$ evolves, we can again focus on how it changes as we add a covariate $l \notin J$. We can then define the ex-ante value of additional covariates simply as the expected reduction in loss it induces, that is, $\mathcal{V}_B^A(J, J\cup\{l\}, n) := \mathcal{L}_B(J,n) - \mathcal{L}_B(J\cup\{l\},n)$.

Proposition 2. *The following is true:*

$$\begin{aligned}\mathcal{V}_B^A(J, J\cup\{l\}, n) &= \mathbb{E}_\pi[\beta_l]^2 + \mathbb{V}_\pi[\beta_l] - \mathbb{E}_{Q_\pi}[\mathbb{V}_\pi[\beta_l \mid D_{J'}]] \\ &\quad + \sum_{j\in J} \mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right] - \mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right]\right] \qquad (6)\end{aligned}$$

If π is atomless over β,

$$\begin{aligned}\mathbb{V}_\pi[\beta_l] - \mathbb{E}_{Q_\pi}[\mathbb{V}_\pi[\beta_l \mid D_{J'}]] &> 0 \\ \mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right] - \mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right]\right] &> 0.\end{aligned}$$

The value of an additional covariate can be decomposed as the sum of three components: (i) $\mathbb{E}_\pi[\beta_l]^2$, the expected forecasting relevance of the additional covariate, now an input to the forecast; (ii) $\mathbb{V}_\pi[\beta_l] - \mathbb{E}_{Q_\pi}[\mathbb{V}_\pi[\beta_l \mid D_{J'}]]$, the expected learning gain on β_l itself; *iii*) the expected learning gains on parameter β_j of the original model, $\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right] - \mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right]\right]$—the expected learning spill-over. The first is weakly positive, the latter two are strictly so. This result parallels our previous one on ex-post valuations, including the presence of learning spill-overs.

Convexity of $\mathcal{L}_B$. Next, we show that the benefit of information displays a form of convexity: if covariates have similar importance, the value of an additional covariate is increasing in the number of covariates already observed. As this analysis faces the same challenges encountered in Section 3.1—analytical solutions are hard to obtain here because of the lack of conjugate priors. Like before, we focus first on (relatively) large samples and provide simulations for small samples.

Consider first the case in which covariates are ex-ante similar: a prior is *symmetric* if $\mathbb{E}_\pi\left[\beta_l^2\right] = \mathbb{E}_\pi\left[\beta_{l'}^2\right] > 0$ for all l, l'.

Proposition 3. *If π is symmetric, then for any J and $l, l' \notin J$, there exist some $\bar{n} \in \mathbb{N}$ such that $\forall n \geq \bar{n}$, $\mathcal{V}_B^A(J \cup \{l\}, J \cup \{l, l'\}, n) > \mathcal{V}_B^A(J, J \cup \{l\}, n)$.*

This result illustrates a form of *convexity in the value of covariates*, meaning that $-\mathcal{L}_B$ is convex in the size of the model. This follows since the *value of an additional covariate $\mathcal{V}_B^A$ is increasing* in the number of covariates. For an intuition, recall that there are two benefits of including an additional covariate: the direct benefit of learning about it—constant under a symmetric prior—and the learning spillover for other covariates. The proposition shows that, as we consider larger and larger models, learning spillovers increase. Intuitively, it shows that the fact that spillovers apply to more covariates is more important than the fact that, once we have little variance left to explain, spillovers for each covariate are smaller. While the result is stated for large enough samples, simulations show that convexity appears robust in small samples as well; Appendix B.2 shows this in detail.

Interestingly, the convexity in model dimensionality stands in contrast to the well-known concavity in dataset size. A classic result is that the gains from one additional observation

$$\mathcal{L}_B(J, n) - \mathcal{L}_B(J, n+1)$$

are *decreasing* in n: the more observations you have, the less valuable an extra observation is. With symmetric priors, the opposite is true of covariates: The more covariates you have, the *more* valuable an extra observation is. As we will discuss, this has several implications.

Beyond Symmetric Priors. The result above on the convexity of the value of covariates relies on the assumptions that each additional covariate is of similar expected explanatory power. If additional covariates are of decreasing expected importance, this result may fail because two forces are at play: the increase in learning spillovers and the decreasing expected importance. To illustrate, consider a prior where covariates are ordered in (linearly) decreasing expected importance: a prior π is *linear* if $\mathbb{E}_\pi\left[\beta_j^2\right] = \lambda(1 + k - j)/k$ for some $\lambda > 0$ for all j. Naturally, agents will consider adding covariates following their order: an agent with model J first considers adding covariate $|J|+1$, then $|J|+2$, and so on. We can then use the shorthand $\mathcal{V}_B^A(|J|, n) := \mathcal{V}_B^A(J, J \cup |J|+1, n)$, since $(|J|, n)$ holds all the information necessary to calculate the value of an additional covariate.

Proposition 4. *If π is linear, then:*

1. *For any k, $\exists \bar{n} > 2k-1$ such that $\forall n \geq \bar{n}$, $\mathcal{V}_B^A(|J|, n)$ is strictly decreasing in $|J|$;*
2. *$\exists \bar{n}$ such that $\forall n \geq \bar{n}$, $\exists k$ such that $\mathcal{V}_B^A(|J|, n)$ is strictly increasing in $|J|$ for $|J| < \frac{2k-n}{3} - 3$, and strictly decreasing for some $|J| \leq k$.*

When observations are abundant relative to covariates and the prior is linear, the value of an extra covariate is always decreasing. This is because with abundant data, learning spillovers are small. But when covariates are more abundant relative to observations, the value of an extra covariate is first increasing and then decreasing. That is, in this case, the value of covariates is first convex and then concave.

3.2.1 Implications

The results above have an immediate implication: when choosing a model before observing data, agents face an inherent pull toward either very small or very large models. To see this, consider an agent with symmetric priors who can add each covariate at a fixed cost $\bar{c}$. Our results imply that such an agent will choose either no covariate at all or the *full* model. If the value of adding one covariate exceeds $\bar{c}$, then the value of each subsequent one must also exceed $\bar{c}$; hence, once the agent starts including covariates, they will never stop. Naturally, the same logic applies if costs decrease with the number of covariates. Consequently, the typical "intermediate" model that includes some but not all covariates can arise only under asymmetric priors or increasing costs. And even in these cases, the same underlying force remains: a bias toward larger models.

Moreover, recall that the convexity in the value of covariates stems from learning spillovers. As these spillovers vanish when n grows large, the tendency to select particularly large models is strongest when samples are not too large—when predictions are less precise and additional covariates are most valuable. As data accumulate further, spillovers become smaller and agents stop buying new covariates, settling either in their small or big models that they decided early on.

4 A Frequentist Approach

While in economic theory it is standard to consider Bayesian agents, forecasts in the real world are often conducted using OLS estimators—a purely frequentist approach.

4.1 A Frequentist Revisiting the Model After Data

Let $\hat{y}_f(x_{Jn})$ be the OLS prediction of y_n by a frequentist who observes the dataset D_J, that is, $\hat{y}_f(x_{Jn}) := x_{Jn}\hat{\beta}_J$ where $\hat{\beta}_J := \left(X_J'X_J\right)^{-1}X_J'Y$, taking $X_J \in \mathbb{R}^{n\times|J|}$ as the matrix form of the observed covariates $\{x_{Ji}\}_{i=0}^{n-1}$ and Y as the analogous vector form of $\{y_i\}_{i=0}^{n-1}$. Note that the distribution of the OLS estimator is given by $\hat{\beta}_J \mid X_J \sim \mathcal{N}\left(\beta_J, \sigma_J^2\left(X_J'X_J\right)^{-1}\right)$. Denote this distribution by $Q_{\hat{\beta}}$.

Expected Loss for a Frequentist. Our first step is to define a frequentist analogue to the Bayesian's expected posterior loss. Consider the following loss function:

$$\begin{aligned}\ell_f\left(\theta_J, D_J\right) &:= \mathbb{E}_{Q_{\theta_J}}\left[y_n - \hat{y}_f(x_{Jn})\right] \\ &= \left(\beta_J - \hat{\beta}_J\left(D_J\right)\right)'\left(\beta_J - \hat{\beta}_J\left(D_J\right)\right) + \sigma_J^2. \end{aligned} \tag{7}$$

This is the expected loss of using OLS to forecast, given all parameters θ_J, where the second line follows from the same steps as in the proof of Lemma 1. Intuitively, the agent's forecast will err when (i) β is not precisely estimated by OLS, and (ii) many important determinants of y are not observed and thus σ_J^2 is high. However, coherent with the frequentist approach, we should also assume that this agent estimates σ_J^2 with frequentist methods, analogous to the treatment of β_J. This gives us the following loss function:

$$\hat{\ell}_f\left(\theta_J, D_J\right) := \left(\beta_J - \hat{\beta}_J\left(D_J\right)\right)'\left(\beta_J - \hat{\beta}_J\left(D_J\right)\right) + s_J^2\left(D_J\right) \tag{8}$$

$$s_J^2\left(D_J\right) := \sum_{i=0}^{n-1} \frac{\left(y_i - \hat{\beta}_J x_{Ji}\right)^2}{n - |J|}. \tag{9}$$

In this case, the distribution of the t-statistic follows $\frac{\hat{\beta}_j - \beta_j}{\sqrt{s_J^2(D_J)\left(X_J'X_J\right)^{-1}_{jj}}} \mid D_J \sim t_{n-|J|}$, where Z_{jj} refers to element (j, j) of matrix Z. Using this distribution, we can integrate over $\hat{\beta}_J$ and note that θ_J drops out of the expression, giving us a definition for the frequentist-equivalent of expected posterior loss,

$$L_f\left(D_J\right) := \mathbb{E}_{Q_{\hat{\beta}}}\left[\hat{\ell}_f\left(\theta_J, D_J\right)\right] = \left(1 + \left(\frac{n - |J|}{n - |J| - 2}\right)\mathrm{tr}\left[\left(X_J'X_J\right)^{-1}\right]\right) s_J^2\left(D_J\right). \tag{10}$$

Expected Loss from the Enlarged Dataset. Next, we want to compute the expected loss from the larger dataset, $\mathbb{E}_{Q_x}\left[L_f\left(D_{J'}\right) \mid D_J\right]$. However, the expected residual variance with an additional covariate is not well defined in standard frequentist statistics, forcing us to make an assumption about how the agent forms beliefs. We consider the case of symmetry across all covariates outside the model,[6] meaning the agent expects a proportional reduction in residual variance, as formalized by the following assumption about how the frequentist forms beliefs.

Assumption 2. *For a frequentist agent considering a larger model $J' = J \cup \{l\}$, the following is true:*

$$\mathbb{E}_{Q_x}\left[s_{J'}^2\left(D_{J'}\right) \mid D_J\right] = \left(\frac{k - |J| - 1}{k - |J|}\right) s_J^2\left(D_J\right).$$

[6]Our results easily extend to non-symmetric beliefs, but that is more cumbersome in notation and provides little additional insight. Therefore, we focus on the symmetric case for exposition.

The Value of an Additional Covariate. Finally, we can define the value of an additional covariate for a frequentist in a way analogous to the one used for Bayesians:

$$\mathcal{V}_f\left(D_J, l\right) := L_f\left(D_J\right) - \mathbb{E}_{Q_x}\left[L_f\left(D_{J'}\right) \mid D_J\right].$$

It turns out that, also for a frequentist, this admits a characterization related to model fit. Adopting the usual frequentist definition of R^2, we obtain the following.

Lemma 3. *Under Assumption 2,*

$$\mathcal{V}_f\left(D_J, l\right) = \frac{g\left(X_J\right)}{1 - |J|/n}\left(1 - R^2\right) s_y^2\left(Y\right)$$

where

$$\begin{aligned} g\left(X_J\right) &:= 1 + \left(\frac{n - |J|}{n - |J| - 2}\right) tr\left[\left(X_J' X_J\right)^{-1}\right] \\ &\quad - \left(\frac{k - |J| - 1}{k - |J|}\right)\left(1 + \left(\frac{n - |J| - 1}{n - |J| - 3}\right)\mathbb{E}_{Q_x}\left[tr\left[\left(X_{J'}' X_{J'}\right)^{-1}\right] \mid D_J\right]\right) \\ s_y^2\left(Y\right) &:= \sum_{i=0}^{n-1} \frac{y_i^2}{n}. \end{aligned}$$

As with the Bayesian, a worse model fit increases the value that an agent places on acquiring a new covariate. Notably, this result holds formally also in small samples—for a fixed estimated variance of y[7] and for a fixed $g\left(X_J\right)$. The latter term captures the difference between the degree of model uncertainty with model J and its expected degree with model J', adjusted by the degrees of freedom. A high $g\left(X_J\right)$ arises when the agent expects to learn a lot about $\beta_{J'}$ with the extra covariate—albeit in the frequentist sense of learning.

Figure 2 illustrates this result using simulations. The left panel plots the value of an additional covariate against the coefficient of determination for randomly drawn X and β.[8] While the value of an additional covariate is clearly inversely related to the model's fit, as was the case for the Bayesian, the relation is far from perfect due to variation in $g\left(X_J\right)$ and $s_y^2\left(Y\right)$. In this case, it is the variation in $s_y^2\left(Y\right)$ that drives this: the right panel plots $\mathcal{V}_f\left(D_J, l\right)/s_y^2$, the value of an extra covariate normalized by the MLE estimator of the variance of y_i, showing a much tighter relationship.

[7]This is the maximum likelihood estimator of the variance when $\mathbb{E}\left[y_i\right] = 0$, since the y_i are normally distributed. However, our takeaways are the same if using the unbiased sample variance, which just rescales by $n/(n-1)$.

[8]Here, X is drawn from the standard normal distribution as specified in Section 2, and the expectation is computed numerically by simulating from that distribution. The vector β is drawn from a distribution that guarantees $\mathbb{V}\left[y_i\right] = 1$ to normalize the scale of the problem. Appendix B.3 has details for the analogous simulation for the Bayesian agent; we follow the same sampling procedure for the frequentist. To compare with the simulation for a Bayesian, we illustrate with $k = 3$, $|J| = 1$ and $n = 10$. We simulate and plot 1000 different samples of X and β.

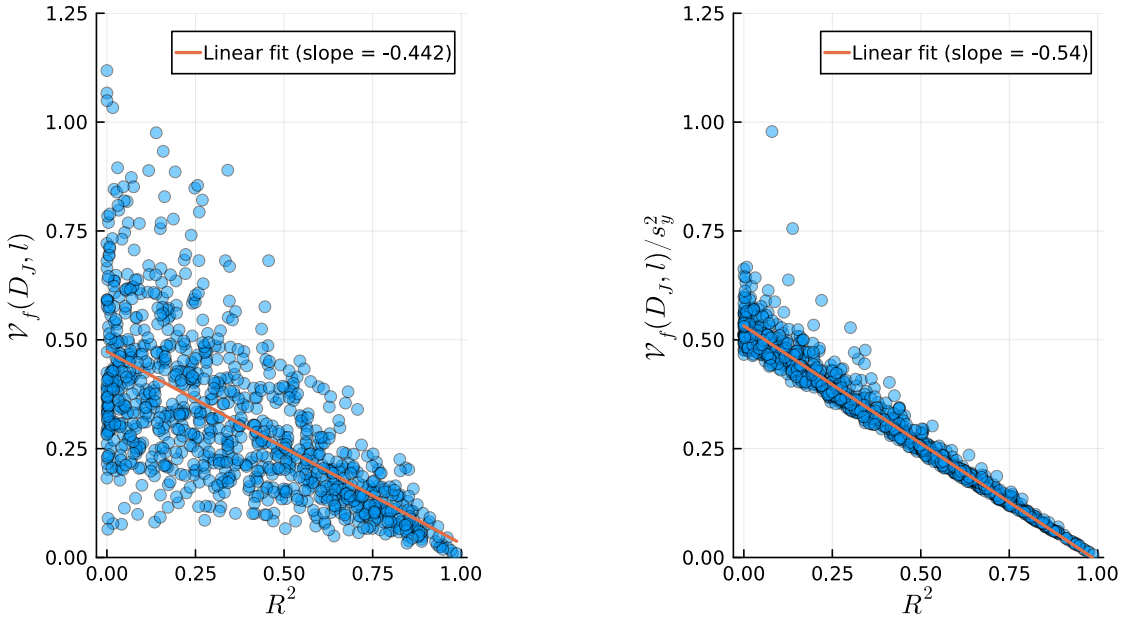


Figure 2: The value of an additional covariate and model fit, frequentist.

Finally, we can consider the asymptotic version of Lemma 3, which (unsurprisingly) converges to the same value as the Bayesian with symmetric beliefs (equal α_j), and so is inversely related to model fit and number of unobserved covariates.

Proposition 5. *Under Assumption 2 and the conditions of Proposition 1,*

$$\mathcal{V}_f\left(D_J, J'\right) \xrightarrow{p} \frac{1}{k - |J|}\left(1 - R^2_\infty\left(J\right)\right)\mathbb{V}\left[y_i\right].$$

Taken together, these results show that the high-level messages we found in the choice to revise a model for a Bayesian continue to hold also for a frequentist, showing the robustness of our core conclusions. As we will see, however, this will not be the case for the choice of the model ex-ante.

4.2 A Frequentist Choosing the Model Ex-Ante

We now consider the model choice J before observing any data, for an agent who knows they will make predictions using OLS. This could be the case, for example, of a company that is deciding which datasets to make available to its employees, knowing that they will use simple regressions in their analysis. As before, for this frequentist to evaluate the benefits of future data, they need to assess the likelihood, and we assume they do so using a prior π. A central object in their decision-making process is

$$\mathcal{L}_f\left(J, n\right) := \mathbb{E}_\pi\left[\mathbb{E}_{Q_{\hat{\beta}}, Q_{X_J}}\left[\ell_f\left(\theta_J, D_J\right)\right]\right],$$

where Q_{X_J} is the exogenous distribution of X_J in the DGP. The inner expectation is similar to L_f defined in Eq. (10), except now we consider the case where expectations over σ_J^2 are formed with π instead of using a frequentist estimate. We can think of $\mathcal{L}_f(J,n)$ as the frequentist analogue of the Bayesian prior loss defined in Eq. (5): the expected loss before observing any data, given the beliefs of an agent who will use frequentist methods. Again, this admits a simple characterization.

Proposition 6. *If $|J| < n-1$, then*

$$\mathcal{L}_f(J,n) := \left(1 + \frac{|J|}{n-|J|-1}\right) \mathbb{E}_\pi\left[\sigma_J^2\right]. \tag{11}$$

For a frequentist, a larger model J has two effects. First, it introduces relevant information for both forecasting and estimating the unknown parameters, reducing σ_J^2. Second, however, it also increases the number of parameters to be estimated, reducing the degrees of freedom $n-|J|-1$ and, overall, the model uncertainty term in parentheses. We can then compute the value of an additional covariate:

$$\begin{aligned}\mathcal{V}_f^A(J, J\cup\{l\}, n) &:= \mathcal{L}_f(J,n) - \mathcal{L}_f(J\cup\{l\},n) \\ &= \left(1 + \frac{|J|}{n-|J|-1}\right)\left(\mathbb{E}_\pi\left[\beta_l^2\right] - \frac{\mathbb{E}_\pi\left[\sigma^2_{J\cup\{l\}}\right]}{n-|J|-2}\right).\end{aligned}$$

A key observation is that, for a frequentist, this expression can be *negative*—this will be the case, if

$$(n-|J|-2)\,\mathbb{E}_\pi\left[\beta_l^2\right] < \mathbb{E}_\pi\left[\sigma^2_{J\cup\{l\}}\right]. \tag{12}$$

In this case, a frequentist will reject an additional covariate, even if it is available for free. This occurs when n is small relative to $|J|$, conditional on the variance in the system $\left(\mathbb{E}_\pi\left[\sigma^2_{J\cup\{l\}}\right]\right)$ and that of the variables to be estimated $\left(\mathbb{E}_\pi\left[\beta_l^2\right]\right)$.

As with the Bayesian, we can obtain additional results by focusing on a symmetric prior.

Proposition 7. *Suppose π is symmetric and $|J| < n-1$. Then,*

$$n \geq k+1 \quad \Leftrightarrow \quad \mathcal{V}_f^A(|J|,n) \geq 0 \quad \Leftrightarrow \quad \Delta_{|J|}\mathcal{V}_f^A(|J|,n) \geq 0.$$

When n is large enough—above $k+1$—the value of an extra covariate is positive *and* increasing in the number of covariates; in this case, the frequentist benefits from more covariates, and increasingly so because increasing the number of covariates allows for a more precise estimation of the original ones. This result mirrors the one we obtained for the Bayesian. However, if n is small, $\mathcal{V}_f^A$ is *negative* and *decreasing*. First, this shows that even with constant priors, extra covariates can

have negative values. Second, it shows that learning spillovers have the *same sign* as the value of an extra covariate: either $\mathcal{V}_f^A$ is positive and increasing, or it is negative and decreasing. Intuitively, when an additional covariate helps estimate, it also provides learning spillovers. When, instead, there are many covariates relative to the number of observations, the loss in degrees of freedom caused by an extra covariate is too costly and therefore harmful for any model size.

We can also study the behavior of the frequentist in the case of a linear prior, mirroring Proposition 4. We find an interesting contrast with the Bayesian.

Proposition 8. *If π is linear, then $\Delta_{|J|}\mathcal{V}_f^A(|J|, n) \geq 0$ for all k, $|J|$ and n.*

This shows that the value of covariates is always concave for a frequentist when the prior is of decreasing importance, in the functional form we have specified. Recall how, instead, a Bayesian with the same prior exhibits, for the initial range, a convexity in the value of covariates because of the strength of the learning spillovers. These spillovers are high for low n, high k, and low $|J|$, the conditions under which the frequentist is "hungry" for degrees of freedom, making ever larger models less attractive.

A "Sophisticated" Frequentist Approach. The frequentist described above may look exceedingly naive—if an extra covariate has a negative value, can't they simply ignore it? In general, shouldn't we allow the agent to ignore some covariates if it chooses? While a Bayesian would generically never choose to do so, a frequentist may. We call such an agent a *sophisticated frequentist*: they will use a frequentist estimate as above, but will choose which variables to consider from the available ones. Notably, we allow this choice to be made *after* the realization of the data.

Since a frequentist using model J has expected loss after the data of $\sigma_J^2\left(1+\text{tr}\left[\left(X_J'X_J\right)^{-1}\right]\right)$, then the expected loss for a sophisticated frequentist who has the covariates in $\tilde{J}$ available is

$$\mathcal{L}_s\left(\tilde{J}, n\right) := \mathbb{E}_{Q_\pi}\left[\min_{J\subseteq\tilde{J}:|J|<n-1} \mathbb{E}_\pi\left[\sigma_J^2\right]\left(1+\text{tr}\left[\left(X_J'X_J\right)^{-1}\right]\right)\right]. \tag{13}$$

Assume also $|J| < n-1$, so that agents have at least 1 degree of freedom with model J and the expected loss can be computed.

While a sophisticated frequentist will never have negative values for more covariates, they will tend to select for smaller models when there are many parameters relative to the number of observations. We can formalize this claim for a single covariate under symmetric priors. Denote the J chosen in the minimization as J^*.

Proposition 9. *If π is symmetric and $|\tilde{J}| = 1$,*

$$\mathbb{P}\left(J^* = \emptyset\right) = F_n\left(k-1\right)$$

Table 1: Probability of no acquisition of covariates, $|\tilde{J}| = 4$

$\mathbb{P}(J^* = \emptyset)$	$k = 5$	$k = 10$	$k = 15$	$k = 20$	$k = 25$
$n = 5$	0.018	0.625	0.940	0.992	0.999
$n = 10$	0.000	0.038	0.461	0.847	0.969
$n = 15$	0.000	0.000	0.044	0.376	0.766
$n = 20$	0.000	0.000	0.001	0.048	0.325
$n = 25$	0.000	0.000	0.000	0.001	0.049

where $\mathbb{P}(A)$ denotes the probability of event A and $F_n(\cdot)$ is the cumulative distribution function of a χ^2 random variable with n degrees of freedom.

That is, a sophisticated frequentist tends to give up on incorporating even the first covariate when there are many parameters relative to observations. Table 1 shows the results of simulations for the case of $|\tilde{J}| > 1$, again for symmetric priors: it shows the probability of giving up completely for different values of k and n when $|\tilde{J}| = 4$. For any n, as k grows large, this probability increases, and it is very high for low values of n.

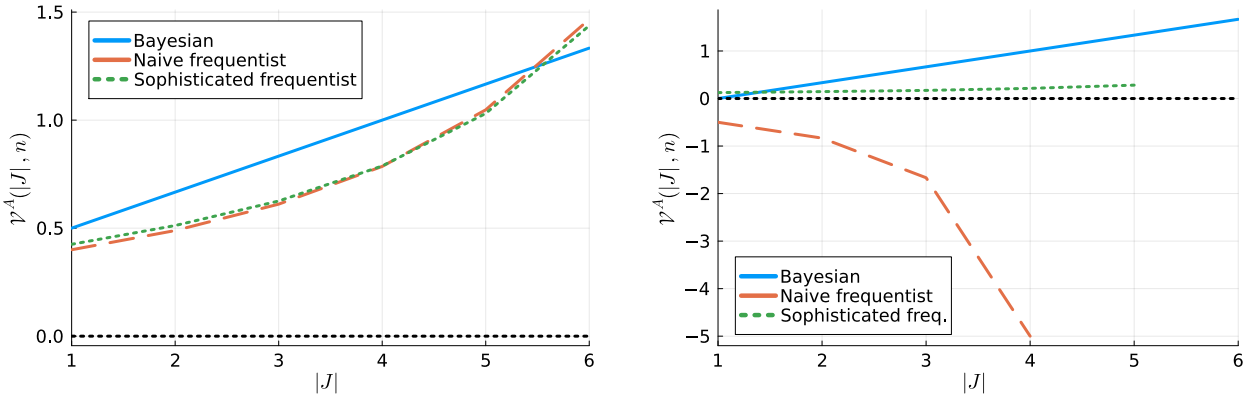
Comparing Bayesians, Frequentists, and Sophisticated Frequentists. To illustrate the differences in the ex-ante valuation between Bayesians,[9] frequentists, and sophisticated frequentists, Figure 3 plots it for the case of $n = 12$, in Panel (a), and $n = 6$, in Panel (b), when $k = 7$ and the prior is symmetric with $\mathbb{E}_\pi\left[\beta_j^2\right] = 1$ for all j. When $n = 12$, we have sufficient degrees of freedom, and a positive and increasing value for all approaches. Moreover, valuations of both types of frequentists are close, since the sophisticated frequentist typically uses all available data. When $n = 6$, however, the naive estimate has a negative value, and this becomes increasingly negative as we add covariates. The value is positive for the sophisticated frequentist, but only slightly increasing, due to the relatively small learning spillovers across covariates. In contrast, the value for Bayesian is clearly positive and increasing.

5 Discussion and Conclusions

Correlated Covariates. While we assumed independent covariates for simplicity, our qualitative insights extend to correlated regressors, with some additional subtleties. First of all, any correlation between observed and unobserved covariates will be exploited by the agents in their forecast, because then the observables carry more predictive power than in the independent case.[10] This implies that, if there is a lot of correlation between observables and unobservable covariates, learning

[9]We use the large samples approximation for the Bayesian, outlined in the proof to Proposition 3. The takeaway from this figure is unchanged if computing the small samples measures.

[10]Note that in our pure prediction problem, agents are not interested in disentangling causal effects across covariates.



(a) $n = 12$ (b) $n = 6$

Figure 3: The value of an additional covariate with a symmetric prior: $k = 7$ and a symmetric prior with $\mathbb{E}_\pi\left[\beta_j^2\right] = 1$ for all j.

spillovers are dampened because of a smaller reduction in noise, since the correlation reduced it to begin with. The opposite holds when the correlation between unobservable covariates is high: in that case, including a new covariate is informative about more than just itself; it's informative about all the remaining unobservable covariates, and learning spillovers are amplified. These additional forces make the trade-off between direct and indirect effects more complex.

Discussion. We examined how decision-makers optimally choose the extent of their own misspecification: agents who recognize that their models omit potentially relevant covariates but face a cost of expanding them. Within a canonical linear-regression environment, we characterized the value of adding an additional covariate under both Bayesian and frequentist learning, and under two different timings.

For Bayesians, two forces shape the incentives to enlarge a model: the direct value of the new covariate—its expected explanatory power—and the indirect learning spillovers that arise because better estimation of one coefficient improves inference about the others. When choosing ex-ante, we show that these spillovers generate a form of convexity in the ex-ante value of covariates: the expected benefit of an additional covariate increases with the number already included. This stands in contrast with the well-known concavity in the value of additional observations. Consequently, when model size is chosen ex ante under symmetric priors and non-growing costs, only corner solutions are attained: either no covariates or all of them; intermediate models arise only when priors are asymmetric or when the cost of expansion grows sufficiently fast.

The frequentist inherits some but not all of these results. However, the ex-ante valuation for a frequentist can be negative when the number of covariates is high relative to the number of observations—the loss of degrees of freedom outweighs the informational gain. Moreover, when

they believe covariates are decreasing linearly in importance, they always have a concave valuation of covariates, as opposed to the Bayesian, who exhibits convex valuation for the initial range. A "sophisticated" frequentist who can ignore redundant variables avoids negative values, yet still displays much flatter incentives, with smaller and less persistent spillovers across covariates.

When decisions are taken after observing data, our results show that the endogenous selection of model richness generates systematic patterns in forecasting and behavior. Agents expand their models when their current specification fits poorly, when few alternative covariates remain, or when theoretical or empirical developments shift their beliefs about excluded variables, reducing uncertainty about which are most relevant. As a result, fully Bayesian agents—as well as frequentists—may display abrupt "jumps" in predictions that can be mistaken for violations of Bayesian updating, offering a fully Bayesian microfoundation for apparently non-Bayesian behavior.

Appendix

A Proofs

A.1 Proof of Lemma 1

Define the Bayesian loss function as the expected loss over realizations of the data, taken parameters and the Bayesian optimal forecast as given.

$$\begin{aligned}
\ell_B(\beta) &:= \mathbb{E}_{Q_\theta}\left[(y_n - \hat{y}_B(x_{Jn}))^2\right] \\
&= \mathbb{E}_{Q_\theta}\left[(x_{Jn}\beta_J + x_{-Jn}\beta_{-J} - x_{Jn}\mathbb{E}_\pi\left[\beta_J \mid D_J\right])^2\right] \\
&= \mathbb{E}_{Q_\theta}\left[(x_{Jn}\left(\beta_J - x_{Jn}\mathbb{E}_\pi\left[\beta_J \mid D_J\right]\right) + x_{-Jn}\beta_{-J})^2\right] \\
&= \sum_{j\in J}\left(\beta_j - \mathbb{E}_\pi\left[\beta_j \mid D_J\right]\right)^2 + \sum_{j\notin J}\beta_j^2.
\end{aligned}$$

The second equality plugs in the optimal forecast into the loss function and uses the definition of y_n. The fourth equality follows from the independence of x_{jn} across j, in addition to their mean of zero and variance of unity. Now, taking the expectation over β with the posterior distribution:

$$\begin{aligned}
L_B(D_J) &= \mathbb{E}_\pi\left[\ell_B(\beta) \mid D_J\right] \\
&= \mathbb{E}_\pi\left[\sum_{j\in J}\left(\beta_j - \mathbb{E}_\pi\left[\beta_j \mid D_J\right]\right)^2 + \sum_{j\notin J}\beta_j^2 \mid D_J\right] \\
&= \sum_{j\in J}\mathbb{V}_\pi\left[\beta_j \mid D_J\right] + \sum_{j\notin J}\mathbb{E}_\pi\left[\beta_j^2 \mid D_J\right].
\end{aligned}$$

□

A.2 Proof of Lemma 2

Using (2) and the definition (3), and recalling $J' = J \cup \{l\}$, the gain of incorporating covariate l (having observed D_J) is given by:

$$\begin{aligned}
\mathcal{V}_B(D_J, l) &= \sum_{j\in J}\mathbb{V}_\pi\left[\beta_j \mid D_J\right] + \sum_{j\notin J}\mathbb{E}_\pi\left[\beta_j^2 \mid D_J\right] - \mathbb{E}_{Q_x}\left[\sum_{j\in J'}\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right] + \sum_{j\notin J'}\mathbb{E}_\pi\left[\beta_j^2 \mid D_{J'}\right] \mid D_J\right] \\
&= \sum_{j\in J}\mathbb{V}_\pi\left[\beta_j \mid D_J\right] + \sum_{j\notin J}\mathbb{E}_\pi\left[\beta_j^2 \mid D_J\right] - \sum_{j\in J'}\mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right] \mid D_J\right] - \sum_{j\notin J'}\mathbb{E}_\pi\left[\beta_j^2 \mid D_J\right],
\end{aligned}$$

where the second line uses the linearity of expectations and the law of total expectation. Rearranging and canceling terms across the summations:

$$\begin{aligned}\mathcal{V}_B\left(D_J, l\right) &= \mathbb{E}_\pi\left[\beta_l^2 \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_l \mid D_{J'}\right] \mid D_J\right] \\ &\quad + \sum_{j \in J}\left(\mathbb{V}_\pi\left[\beta_j \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right) \\ &= \mathbb{E}_\pi\left[\beta_l \mid D_J\right]^2 + \mathbb{V}_\pi\left[\beta_l \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_l \mid D_{J'}\right] \mid D_J\right] \\ &\quad + \sum_{j \in J}\left(\mathbb{V}_\pi\left[\beta_j \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right),\end{aligned}$$

where the second line uses the definition of variance. Rearranging gives the expression in the lemma.

□

A.3 Proof of Proposition 1

Doob's consistency theorem implies that the posterior converges in distribution inside $\tilde{\Theta}_J$, having no mass elsewhere. That is, every combination of the parameters that does not satisfy the simplex is ruled out, because it is not consistent with the variance of the residuals σ_J^2 that the agent has learned with certainty. Therefore, when characterizing the posterior we should only care about its distribution within the simplex $\tilde{\Theta}_J$ characterized by the true value of σ_J^2. Further, since the distribution within this simplex is not identified, the posterior retains the shape of the prior. Since the posterior is jointly Dirichlet-distributed, the marginal posterior distributions are Beta-distributed. Let $\alpha_0 := \sum_{j \notin J} \alpha_j$, then:

$$\tilde{\beta}_j^2 \mid D_J \sim \text{Beta}\left(\alpha_j, \alpha_0 - \alpha_j\right) \quad \forall j \notin J.$$

The implication is that the posterior mean for the normalized and unnormalized parameters follow:

$$\begin{aligned}\mathbb{E}_\pi\left[\tilde{\beta}_j^2 \mid D_J\right] &= \frac{\alpha_j}{\alpha_0}, \\ \mathbb{E}_\pi\left[\beta_j^2 \mid D_J\right] &= \frac{\alpha_j}{\alpha_0}\sigma_J^2,\end{aligned} \tag{14}$$

where the second line uses the definition $\tilde{\beta}_j := \beta_j/\sigma_J$. Now, assumption 1 directly implies that for all $j \in J$, as $n \to \infty$:

$$\mathbb{V}_\pi\left[\beta_j \mid D_J\right] \xrightarrow{p} 0.$$

Intuitively there is no uncertainty about the regression coefficients inside the model in the limit. We can use this result in (4) to conclude that:

$$\mathcal{V}_B\left(D_J, l\right) \xrightarrow{p} \mathbb{E}_\pi\left[\beta_l \mid D_J\right]^2 + \mathbb{V}_\pi\left[\beta_l \mid D_J\right] = \mathbb{E}_\pi\left[\beta_l^2 \mid D_J\right].$$

Plugging (14) into this last expression finishes the proof.

□

A.4 Proof of Proposition 2

The expression $\mathcal{V}_B^A\left(J, J \cup \{l\}, n\right)$ can be obtained by using its definition and (5), as well as the definition of variance $\mathbb{V}_\pi[\beta_l] = \mathbb{E}_\pi[\beta_l^2] - \mathbb{E}_\pi[\beta_l]^2$. The first inequality is a direct application of the law of total variance:

$$\mathbb{V}_\pi\left[\beta_l\right] - \mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_l \mid D_{J'}\right]\right] = \mathbb{V}_{Q_\pi}\left[\mathbb{E}_\pi\left[\beta_l \mid D_{J'}\right]\right] \geq 0$$

This expression is nonnegative because it is a second moment, and now we show that it is strictly positive. With some abuse of notation, let $\pi\left(\beta\right)$ denote the prior density and $\pi\left(\beta \mid D_{J'}\right)$ the posterior density. Consider some β with positive density under the prior. From Bayes' law:

$$\pi\left(\beta \mid D_{J'}\right) \propto \left(\sum_{j \notin J'} \beta_j^2\right)^{-n/2} \exp\left(-\frac{\sum_{i=0}^{n-1}\left(y_i - \sum_{j \in J'} x_{ij}\beta_j\right)^2}{2\sum_{j \notin J'} \beta_j^2}\right) \pi\left(\beta\right)$$

Note that all elements all elements in this expression are strictly positive. Now consider a perturbation in $D_{J'}$: y_ℓ increases in observation $\ell \in \{0, \ldots, n-1\}$. Since π is atomless, the posterior must follow:

$$\frac{\partial \pi\left(\beta \mid D_{J'}\right)}{\partial y_\ell} \propto \left(\sum_{j \notin J'} \beta_j^2\right)^{-n/2} \exp\left(-\frac{\sum_{i=0}^{n-1}\left(y_i - \sum_{j \in J'} x_{ij}\beta_j\right)^2}{2\sum_{j \notin J'} \beta_j^2}\right) \pi\left(\beta\right) \left(\frac{\sum_{j \in J'} x_{\ell j}\beta_j - y_\ell}{\sum_{j \notin J'} \beta_j^2}\right)$$

Clearly, the posterior density satisfies:

$$\begin{aligned}
\frac{\partial \pi\left(\beta \mid D_{J'}\right)}{\partial y_\ell} > 0 \quad &\iff \quad y_\ell < \sum_{j \in J'} x_{\ell j}\beta_j \\
&\iff \quad \beta_l \begin{cases} > \frac{y_\ell - \sum_{j \in J', j \neq \ell} x_{\ell j}\beta_j}{x_{\ell l}} & \text{if } x_{\ell l} > 0 \\ < \frac{y_\ell - \sum_{j \in J', j \neq \ell} x_{\ell j}\beta_j}{x_{\ell l}} & \text{if } x_{\ell l} < 0 \end{cases}
\end{aligned}$$

Conditional on the data and all other parameters $\beta \backslash \beta_l$, this expression determines a threshold for β_l that determines whether the posterior density of the joint distribution of β increases or decreases.

If the conditional density of $\beta_l \mid \beta \backslash \beta_l, D_J$ satisfies this threshold rule, then a higher realization of y_ℓ raises (lowers) $\mathbb{E}_\pi \left[\beta_l \mid \beta \backslash \beta_l, D_{J'}\right]$ when $x_{\ell l}$ is positive (negative). Since this last statement holds for all $\beta \backslash \beta_l$, then it also holds for the marginal density of $\beta_l \mid D_J$. We conclude that $\mathbb{E}_\pi \left[\beta_l \mid D_{J'}\right]$ is increasing (decreasing) in y_ℓ when when $x_{\ell l}$ is positive (negative). Lastly, because y_ℓ has a continuous positive density over $\mathbb{R}$, it must be that $\mathbb{V}_{Q_\pi}\left[\mathbb{E}_\pi \left[\beta_l \mid D_{J'}\right]\right] > 0$. This establishes the first inequality.

For the second inequality, first expand it through the law of total expectations:

$$\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi \left[\beta_j \mid D_J\right] - \mathbb{V}_\pi \left[\beta_j \mid D_{J'}\right]\right] = \mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi \left[\beta_j \mid D_J\right] - \mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi \left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right] \tag{15}$$

Where the outer expectation is over D_J, and the inner expectation is over $D_{J'} \backslash D_J = \{x_{li}\}_{i=0}^{n-1}$. By the law of total variance:

$$\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi \left[\beta_j \mid D_{J'}\right] \mid D_J\right] = \mathbb{V}_\pi \left[\beta_j \mid D_J\right] - \mathbb{V}_{Q_\pi}\left[\mathbb{E}_\pi \left[\beta_j \mid D_{J'}\right] \mid D_J\right]$$

Plugging this expression into (15):

$$\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi \left[\beta_j \mid D_J\right] - \mathbb{V}_\pi \left[\beta_j \mid D_{J'}\right]\right] = \mathbb{E}_{Q_\pi}\left[\mathbb{V}_{Q_\pi}\left[\mathbb{E}_\pi \left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right] \geq 0$$

This expression is strictly positive under the same argument as with the first inequality.

□

A.5 Proof of Proposition 3

Using the asymptotic approximation (1) in the definition of Bayesian prior loss given by (5), we obtain:

$$\begin{aligned}
\mathcal{L}_B(J, n) &= \left(\frac{|J|}{n}\right) \mathbb{E}_\pi \left[\sigma_J^2\right] + \sum_{j \notin J} \mathbb{E}_\pi \left[\beta_j^2\right] + o\left(n^{-1}\right) \\
&= \left(1 + \frac{|J|}{n}\right) \mathbb{E}_\pi \left[\sigma_J^2\right] + o\left(n^{-1}\right)
\end{aligned}$$

Where we have used the definition of σ_J^2 in the second line. A prior is symmetric when $\mathbb{E}_\pi[\beta_j^2] = \lambda$ for all j, for some $\lambda > 0$. Then, the asymptotic approximation of the Bayesian prior loss with a symmetric prior follows:

$$\mathcal{L}_B\left(J, n\right) = \left(1 + \frac{|J|}{n}\right) \lambda \left(k - |J|\right) + o\left(n^{-1}\right)$$

The value of another covariate is then given by:

$$\begin{aligned}\mathcal{V}_B^A\left(|J|,n\right) &= \mathcal{L}_B\left(J,n\right)-\mathcal{L}_B\left(J\cup\{l\},n\right)\\ &= \lambda\left(1+\frac{|J|+1}{n}\right)-\frac{\lambda\left(k-|J|\right)}{n}+o\left(n^{-1}\right)\end{aligned}$$

Note that because of the symmetric prior assumption, this value now depends only on the number of covariates in the model, instead of the specific model and covariate to be added. With some abuse of notation, we now use the number of covariates as an argument to this function. Then, the first difference of this value is given by:

$$\begin{aligned}\Delta_{|J|}\mathcal{V}_B^A\left(|J|,n\right) &:= \mathcal{V}_B^A\left(|J|+1,n\right)-\mathcal{V}_B^A\left(|J|,n\right)\\ &= \frac{2\lambda}{n}+o\left(n^{-1}\right)\end{aligned}$$

Denote the remainder term as $r\left(n\right)=o\left(n^{-1}\right)$. By its order, there exist an $\bar{n}$ such that for every $n\geq\bar{n}$, we have:

$$\left|r\left(n\right)\right|\leq\frac{\lambda}{n}$$

For these values of n:

$$\Delta_{|J|}\mathcal{V}_B^A\left(|J|,n\right)=\frac{2\lambda}{n}+r\left(n\right)\geq\frac{2\lambda}{n}-\frac{\lambda}{n}>0$$

□

A.6 Proof of Proposition 4

First, note that the expected residual variance can be expressed as:

$$\begin{aligned}\mathbb{E}_\pi\left[\sigma_J^2\right] &= \lambda\sum_{j=|J|+1}^{k}\frac{1+k-j}{k}\\ &= \frac{\lambda\left(k+k^2-|J|-2k|J|+|J|^2\right)}{2k}\end{aligned}\tag{16}$$

The Bayesian prior loss follows:

$$\begin{aligned}\mathcal{L}_B\left(|J|,n\right) &= \left(1+\frac{|J|}{n}\right)\mathbb{E}_\pi\left[\sigma_J^2\right]+o\left(n^{-1}\right)\\ &= \left(1+\frac{|J|}{n}\right)\frac{\lambda}{2k}\left[k+k^2-\left(1+2k\right)|J|+|J|^2\right]+o\left(n^{-1}\right)\end{aligned}$$

The value of a covariate is given by:

$$\begin{aligned}\mathcal{V}_B^A(|J|,n) &= \mathcal{L}_B(|J|,n) - \mathcal{L}_B(|J|+1,n) \\ &= \left(1+\frac{|J|}{n}\right)\frac{\lambda}{k}(k-|J|) - \frac{\lambda}{2kn}\left[k+k^2-(1+2k)(|J|+1)+(|J|+1)^2\right] + o\left(n^{-1}\right)\end{aligned}$$

Its first difference is:

$$\begin{aligned}\Delta_{|J|}\mathcal{V}_B^A(|J|,n) &= \mathcal{V}_B^A(|J|+1,n) - \mathcal{V}_B^A(|J|,n) + o\left(n^{-1}\right) \\ &= \frac{\lambda}{kn}(2k-3|J|-2-n) + o\left(n^{-1}\right)\end{aligned}$$

For the first statement, denote the remainder as $r(n)$ and note that:

$$\Delta_{|J|}\mathcal{V}_B^A(|J|,n) \leq \Delta_{|J|}\mathcal{V}_B^A(0,n) = \frac{\lambda}{kn}(2k-2-n) + r(n)$$

Since $r(n) = o\left(n^{-1}\right)$, there exist an $\bar{n}$ such that for every $n \geq \bar{n}$, we have:

$$|r(n)| \leq \frac{\lambda}{kn}$$

For these values of n:

$$\Delta_{|J|}\mathcal{V}_B^A(0,n) \leq \frac{\lambda}{kn}(2k-2-n) + \frac{\lambda}{kn} = \frac{\lambda}{kn}(2k-1-n)$$

This last term is negative when $n > 2k-1$, thus proving the first statement for $|J| = 0$. For the second one, using the same bound for the absolute value of the remainder term:

$$\Delta_{|J|}\mathcal{V}_B^A(|J|,n) \geq \frac{\lambda}{kn}(2k-3|J|-2-n) - \frac{\lambda}{kn} = \frac{\lambda}{kn}(2k-3|J|-3-n)$$

The last term is strictly positive when:

$$|J| < \frac{2k-n}{3} - 1$$

As long as $k > (n+3)/2$, then $\Delta_{|J|}\mathcal{V}_B^A(0,n) > 0$, proving the first part of the second statement. For this same n, k, and using the same bound for the remainder term:

$$\Delta_{|J|}\mathcal{V}_B^A(k,n) \leq \frac{\lambda}{kn}(-k-2-n) + \frac{\lambda}{kn} < 0$$

Completing the second part of the second statement for $|J| = k$. □

A.7 Proof of Lemma 3

Starting from the definition of the value of an additional covariate, and using (10):

$$\begin{aligned}\mathcal{V}_f\left(D_J, l\right) &= L_f\left(D_J\right) - \mathbb{E}_{Q_x}\left[L_f\left(D_{J'}\right) \mid D_J\right] \\ &= s_J^2\left(D_J\right)\left(1 + \left(\frac{n-|J|}{n-|J|-2}\right)\mathrm{tr}\left[\left(X_J'X_J\right)^{-1}\right]\right) \\ &\quad -\mathbb{E}_{Q_x}\left[s_{J'}^2\left(D_{J'}\right)\left(1 + \left(\frac{n-|J|-1}{n-|J|-3}\right)\mathrm{tr}\left[\left(X_{J'}'X_{J'}\right)^{-1}\right]\right) \mid D_J\right]\end{aligned}$$

Imposing Assumption 2:

$$\begin{aligned}\mathcal{V}_f\left(D_J, l\right) &= s_J^2\left(D_J\right)\left(1 + \left(\frac{n-|J|}{n-|J|-2}\right)\mathrm{tr}\left[\left(X_J'X_J\right)^{-1}\right]\right) \\ &\quad -s_J^2\left(D_J\right)\left(\frac{k-|J|}{1+k-|J|}\right)\left(1 + \left(\frac{n-|J|-1}{n-|J|-3}\right)\mathbb{E}_{Q_x}\left[\mathrm{tr}\left[\left(X_{J'}'X_{J'}\right)^{-1}\right] \mid D_J\right]\right) \\ &= g\left(X_J\right)s_J^2\left(D_J\right) \qquad (17)\end{aligned}$$

Where $g\left(X_J\right)$ is defined as in Proposition 3. Now, the coefficient of determination for this model is defined as:

$$\begin{aligned}R^2 &:= 1 - \frac{\sum_{i=0}^{n-1}\left(y_i - x_{Ji}\hat{\beta}_J\right)^2}{\sum_{i=0}^{n-1} y_i^2} \\ \sum_{i=0}^{n-1}\left(y_i - x_{Ji}\hat{\beta}_J\right)^2 &= \left(1-R^2\right)\sum_{i=0}^{n-1} y_i^2\end{aligned}$$

Plug that into the definition of $s_J^2\left(D_J\right)$, given by (9):

$$s_J^2\left(D_J\right) = \left(\frac{1}{n-|J|}\right)\left(1-R^2\right)\sum_{i=0}^{n-1} y_i^2$$

Inserting this expression in (17)

$$\mathcal{V}_f\left(D_J, l\right) = \frac{g\left(X_J\right)}{n-|J|}\left(1-R^2\right)\sum_{i=0}^{n-1} y_i^2$$

Multiplying and dividing by n yields the result.

□

A.8 Proof of Proposition 5

Recall that for any finite n we have:

$$\mathcal{V}_f\left(D_J, J'\right) = \frac{g\left(X_J\right)}{1-|J|/n}\left(1-R^2\left(D_J\right)\right)s_y^2\left(Y\right)$$

We analyze the convergence of each of the terms in order. Our first step is to show that $g\left(X_J\right) \xrightarrow{p} 1/(1+k-|J|)$. To this end, note that since X_J follows a multivariate normal distribution, then $\left(X_J'X_J\right)^{-1}$ follows an inverse Wishart distribution with the identity matrix as the scale matrix, and n degrees of freedom. From the properties of the inverse Wishart, direct calculation shows that for any $n > |J|+3$:

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr}\left(\left(X_J'X_J\right)^{-1}\right)\right] &= \frac{|J|}{n-|J|-1} \\
\mathbb{V}\left[\operatorname{tr}\left(\left(X_J'X_J\right)^{-1}\right)\right] &= \frac{2|J|\left(n-1\right)}{\left(n-|J|\right)\left(n-|J|-1\right)^2\left(n-|J|-3\right)}
\end{aligned}$$

Then as $n \to \infty$, both the mean and the variance of this random variable converge to 0. This implies that:

$$\operatorname{tr}\left[\left(X_J'X_J\right)^{-1}\right] \xrightarrow{p} 0$$

From its definition, it then follows that $g\left(X_J\right)$ converges:

$$g\left(X_J\right) \xrightarrow{p} 1 - \frac{k-|J|}{1+k-|J|} = \frac{1}{1+k-|J|} \tag{18}$$

Now we turn to the convergence of $R^2\left(D_J\right)$. First, note that our DGP outlined in section **??** satisfies the necessary assumptions so that $\hat{\beta}_J \xrightarrow{p} \beta_J$.[11] This implies $y_i - x_{Ji}\hat{\beta}_J \xrightarrow{p} y_i - x_{Ji}\beta_J = \varepsilon_{Ji}$, and in turn we conclude:

$$\sum_{i=0}^{n-1}\frac{\left(y_i - x_{Ji}\hat{\beta}_J\right)^2}{n} \xrightarrow{p} \sigma_J^2 = \mathbb{V}\left[\varepsilon_{Ji}\right]$$

Since the above is a consistent estimator of the variance of residuals (and $\mathbb{E}\left[\varepsilon_{Ji}\right] = 0$). The same principle applies to the estimator of the variance of y_i:

$$s_y^2\left(Y\right) \xrightarrow{p} \mathbb{V}\left[y_i\right] \tag{19}$$

[11] For example, from Hayashi (2000): (1) our DGP is linear; (2) the joint distribution of $\{x_i, y_i\}$ is stationary and ergodic; (3) regressors are always orthogonal to errors (which include other covariates) because all covariates are independent between each other; (4) and $\mathbb{E}\left[x_i'x_i\right] = I$ is nonsingular.

Then, using the definition of the coefficient of determination:

$$R^2\left(D_J\right) = 1 - \frac{\sum_{i=0}^{n-1}\left(y_i - x_{Ji}\hat{\beta}_J\right)^2}{\sum_{i=0}^{n-1} y_i^2} = 1 - \frac{\sum_{i=0}^{n-1}\left(y_i - x_{Ji}\hat{\beta}_J\right)^2/n}{s_y^2\left(Y\right)} \xrightarrow{p} 1 - \frac{\sigma_J^2}{\mathbb{V}\left[y_i\right]} \tag{20}$$

Combining (18)-(20) yields the desired result.

□

A.9 Proof of Proposition 6

First, here we reproduce equation (7):

$$\ell_f\left(\theta_J, D_J\right) = \left(\beta_J - \hat{\beta}_J\left(D_J\right)\right)'\left(\beta_J - \hat{\beta}_J\left(D_J\right)\right) + \sigma_J^2.$$

Consider now the frequentist-equivalent of the expected posterior loss, by integrating over $\hat{\beta}_J$:

$$\begin{aligned}\mathbb{E}_{Q_{\hat{\beta}}}\left[\ell_f\left(\theta_J, D_J\right)\right] &= \mathbb{E}_{Q_{\hat{\beta}}}\left[\text{tr}\left[\left(\beta_J - \hat{\beta}_J\right)'\left(\beta_J - \hat{\beta}_J\right)\right]\right] + \sigma_J^2 \\ &= \text{tr}\left[\mathbb{E}_{Q_{\hat{\beta}}}\left[\left(\beta_J - \hat{\beta}_J\right)\left(\beta_J - \hat{\beta}_J\right)'\right]\right] + \sigma_J^2\end{aligned}$$

Where we use the properties of the trace operator. Since the distribution of the OLS estimator is given by $\hat{\beta}_J \mid X_J \sim \mathcal{N}\left(\beta_J, \sigma_J^2\left(X_J'X_J\right)^{-1}\right)$, this expectation satisfies:

$$\mathbb{E}_{Q_{\hat{\beta}}}\left[\ell_f\left(\theta_J, D_J\right)\right] = \text{tr}\left[\sigma_J^2\left(X_J'X_J\right)^{-1}\right] + \sigma_J^2$$

Now integrate over X_J to get:

$$\begin{aligned}\mathbb{E}_{Q_{X_J}, Q_{\hat{\beta}}}\left[\ell_f\left(\theta_J, D_J\right)\right] &= \mathbb{E}_{Q_{X_J}}\left[\text{tr}\left[\sigma_J^2\left(X_J'X_J\right)^{-1}\right]\right] + \sigma_J^2 \\ &= \sigma_J^2\text{tr}\left[\mathbb{E}_{Q_{X_J}}\left[\left(X_J'X_J\right)^{-1}\right]\right] + \sigma_J^2 \\ &= \sigma_J^2\text{tr}\left[\frac{1}{n - |J| - 1} I_{|J|}\right] + \sigma_J^2 \\ &= \sigma_J^2\left(1 + \frac{|J|}{n - |J| - 1}\right)\end{aligned}$$

We've used the fact that if X_J is an $n \times |J|$ matrix of standard normal random variables, then $\left(X_J'X_J\right)^{-1}$ follows an inverse Wishart distribution with the mean given inside the square brackets in the third line.

Finally, integrate over π to obtain the desired result.

□

A.10 Proof of Proposition 7

We use the same definition of a symmetric prior as in Proposition 3: $\mathbb{E}_\pi[\beta_j^2] = \lambda$ for all j and for some $\lambda > 0$.[12] Then, the average risk for the frequentist follows:

$$\begin{aligned} \mathcal{L}_f(|J|, n) &= \left(1 + \frac{|J|}{n - |J| - 1}\right) \lambda (k - |J|) \\ &= (n-1) \frac{\lambda (k - |J|)}{n - |J| - 1} \end{aligned} \tag{21}$$

Let $\tilde{\mathcal{L}}_f(|J|, n) : \mathbb{R}_+ \times \mathbb{N} \to \mathbb{R}_+$ be the continuous function defined by the right-hand side of (21). A useful object to construct easy proofs of will be the (negative) derivatives with respect to $|J|$:

$$-\frac{\partial \tilde{\mathcal{L}}_f(|J|, n)}{\partial |J|} = \frac{(n-1)\lambda (n-k-1)}{(n - |J| - 1)^2} \tag{22}$$

$$-\frac{\partial^2 \tilde{\mathcal{L}}_f(|J|, n)}{\partial |J|^2} = \frac{2(n-1)\lambda (n-k-1)}{(n - |J| - 1)^3} \tag{23}$$

Notice that the sign of (22) and (23) are the same for all $|J|$, and they only depend on $\lambda(n-k-1) - \phi$ once we assume $n > |J| + 1$. Since for given $\lambda(n-k-1) - \phi$ the sign is the same for all $|J|$, this implies that the same sign holds for the discrete differences $-\Delta_{|J|}\mathcal{L}_f(|J|, n)$ and $-\Delta^2_{|J|}\mathcal{L}_f(|J|, n)$.

□

A.11 Proof of Proposition 8

Using the same technique as in Proposition 7, from (11) we can get the condition for concavity. We need to show:

$$-\frac{\partial^2 \tilde{\mathcal{L}}_f(|J|, n)}{\partial |J|^2} < 0 \iff (n - |J| - 1)^2 \frac{\partial^2 \mathbb{E}_\pi\left[\sigma_J^2\right]}{\partial |J|^2} + 2(n - |J| - 1) \frac{\partial^2 \mathbb{E}_\pi\left[\sigma_J^2\right]}{\partial |J|^2} + 2\mathbb{E}_\pi\left[\sigma_J^2\right] > 0$$

[12] As stated, a symmetric prior implies that as k grows large, holding $|J|$ fixed, $\mathbb{E}_\pi\left[\sigma_J^2\right]$ grows large as well, which might be counterintuitive for the comparative statics. To avoid this, an alternative symmetric prior could specify $\mathbb{E}_\pi\left[\beta_j^2\right] = \lambda/k$ for some $\lambda > 0$. It can be shown that Propositions 3 and 7 would continue to hold in this case as well.

The derivatives of the expected residual variance are:

$$\frac{\partial \mathbb{E}_\pi\left[\sigma_J^2\right]}{\partial |J|} = \frac{\lambda\left(2|J| - 1 - 2k\right)}{2k} \tag{24}$$

$$\frac{\partial^2 \mathbb{E}_\pi\left[\sigma_J^2\right]}{\partial |J|^2} = \frac{\lambda}{k} \tag{25}$$

Evaluating the condition for concavity with (24) and (25), we need:

$$\begin{aligned} \frac{\left(n - |J| - 1\right)^2 + \left(n - |J| - 1\right)\left(2|J| - 1 - 2k\right) + k + k^2 - |J| - 2k|J| + |J|^2}{k} &> 0 \\ n^2 - 3n + 2 - 2nk + 3k + k^2 &> 0 \end{aligned}$$

So we need to show:

$$g\left(n, k\right) := n^2 - n\left(3 + 2k\right) + \left(2 + 3k + k^2\right) \geq 0$$

Where we have rearranged the equation to be quadratic in n given k. For any k, if n were continuous then $g\left(n, k\right)$ would be minimized at:

$$n^* = \frac{3 + 2k}{2}$$

However, $n \in \mathbb{N}$. Therefore, we must evaluate the expression at the closest integers. Since k is an integer, the closest integer n from below is given by:

$$\underline{n}^* = n^* - \frac{1}{2} = 1 + k$$

Evaluating:

$$g\left(\underline{n}^*, k\right) = \left(1 + k\right)^2 - \left(1 + k\right)\left(3 + 2k\right) + 2 + 3k + k^2 = 0$$

Consider now the closest integer from above:

$$\bar{n}^* = n^* + \frac{1}{2} = 2 + k$$

Evaluating:

$$g\left(\bar{n}^*, k\right) = \left(2 + k\right)^2 - \left(2 + k\right)\left(3 + 2k\right) + 2 + 3k + k^2 = 0$$

Therefore, for any k, we have $g\left(n, k\right) \geq 0$ for all n. Retracing our steps, this implies that for all $|J|$ and n we have $-\partial^2 \tilde{\mathcal{L}}_f\left(|J|, n\right) / \partial |J|^2 < 0$. This means the discrete difference over integer $|J|$ is strictly negative as well for all $|J| \in \mathbb{N}$.

□

A.12 Proof of Proposition 9

If π is constant and $|\tilde{J}| = 1$, then following (13) a sophisticated frequentist will choose $J = \emptyset$ relative to the alternative of $J = 1$ when:

$$\begin{aligned}
\mathbb{E}_\pi\left[\sigma_\emptyset^2\right] &< \mathbb{E}_\pi\left[\sigma_1^2\right]\left(1 + \text{tr}\left[\left(X_1'X_1\right)^{-1}\right]\right) \\
\lambda k &< \lambda\left(k-1\right)\left(1 + \frac{1}{\sum_{i=0}^{n-1} x_{1i}^2}\right) \\
\sum_{i=0}^{n-1} x_{1i}^2 &< \left(k-1\right)
\end{aligned}$$

The distribution of the LHS of the inequality above is a χ^2 with n degrees of freedom. □

B Simulations

B.1 Large Samples Approximation Accuracy

The purpose of this section is to evaluate the accuracy of the asymptotic approximation suggested by (1):

$$\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right] \approx \mathbb{E}_\pi\left[\sigma_J^2\right]/n$$

For this purpose we will consider a conjugate prior, which simplifies the analysis tremendously by allowing for the analytical computation of the posterior given some dataset. However, we will still need to analyze expectations over realizations of the data computationally, as they do not have a closed form. For an agent with a model J, we consider a prior π under which:

$$\begin{aligned}
\sigma_J^2 &\sim \Gamma^{-1}\left(a, b\right) \\
\beta_J \mid \sigma_J^2 &\sim \mathcal{N}_{|J|}\left(0, cI_{|J|}\sigma_J^2\right)
\end{aligned}$$

Where Γ^{-1} denotes the inverse gamma distribution with parameters $a \geq 1$ and $b > 0$; $I_{|J|}$ denotes the identity matrix of size $|J|$ and $c > 0$. Under this prior

$$\mathbb{E}_\pi\left[\sigma_J^2\right] = \frac{b}{a-1}$$

A well-known result is that this prior is conjugate, and observing a dataset $D_J = \{X_J, Y\}$ results in the following posterior:

$$\begin{aligned}
\beta_J \mid \sigma_J^2 &\sim \mathcal{N}_{|J|}\left(\hat{\mu}, \hat{\Sigma}\sigma_J^2\right) && (26)\\
\hat{\Sigma} &= \left(X_J'X_J + c^{-1}I_{|J|}\right)^{-1} && (27)\\
\hat{\mu} &= \hat{\Sigma}X_J'Y && (28)\\
\sigma_J^2 &\sim \Gamma^{-1}\left(\hat{a}, \hat{b}\right) && (29)\\
\hat{a} &= a + n/2 && (30)\\
\hat{b} &= b + \left(Y'Y - \hat{\mu}'\hat{\Sigma}^{-1}\hat{\mu}\right)/2 && (31)
\end{aligned}$$

From the posterior distributions, we can characterize the marginal variance of the regression coefficients. By the law of total variance:

$$\begin{aligned}
\mathbb{V}_\pi\left[\beta_J \mid D_J\right] &= \mathbb{E}_\pi\left[\mathbb{V}_\pi\left[\beta_J \mid \sigma_J^2, D_J\right] \mid D_J\right] + \mathbb{V}_\pi\left[\mathbb{E}_\pi\left[\beta_J \mid \sigma_J^2, D_J\right] \mid D_J\right]\\
&= \mathbb{E}_\pi\left[\hat{\Sigma}\sigma_J^2 \mid D_J\right] + \mathbb{V}_\pi\left[\hat{\mu} \mid D_J\right]\\
&= \hat{\Sigma}\mathbb{E}_\pi\left[\sigma_J^2 \mid D_J\right]\\
\mathbb{V}_\pi\left[\beta_J \mid D_J\right] &= \frac{\hat{\Sigma}\hat{b}}{\hat{a}-1} && (32)
\end{aligned}$$

The outer moments integrate over σ_J^2, which is why the second term reduces to 0 (the posterior mean of β_J does not depend on σ_J^2). With these results in hand, given any $\{a, b, c, |J|\}$ we can simulate as many datasets as needed with Monte Carlo methods and compute $\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_J \mid D_J\right]\right]$ numerically. In particular, for each n we:

1. Draw a scalar σ_J^2 from $\Gamma^{-1}(a, b)$
2. Draw a vector β_J from $\mathcal{N}_{|J|}\left(0, cI_{|J|}\sigma_J^2\right)$, using the σ_J^2 drawn in the previous step
3. Simulate X_J with n draws from $\mathcal{N}_{|J|}\left(0, I_{|J|}\right)$
4. Simulate ε_J with n draws from $\mathcal{N}\left(0, \sigma_J^2\right)$ and compute Y
5. Obtain the parameters $\left\{\hat{\Sigma}, \hat{a}, \hat{b}\right\}$ following (26)-(31).
6. Compute $\mathbb{V}_\pi\left[\beta_J \mid D_J\right]$ with (32)
7. Repeat steps 1-6 a large number of times and store results
8. Compute $\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right]$ by averaging over all $\mathbb{V}_\pi\left[\beta_l \mid D_J\right]$ for $l \in J$.

Parameters $\{a, b\}$ are found not to matter for our takeaways, other than scaling the size of the problem.[13]. Therefore, we illustrate our results with $a = b + 1$ and the limit $b \to \infty$, which yields $\mathbb{E}_\pi\left[\sigma_J^2\right] = 1$, as well as $\mathbb{V}_\pi\left[\sigma_J^2\right] \to 0$. This limit reduces simulation noise to a minimum. We impose one normalization to keep the scale of the problem constant:

$$\mathbb{V}_\pi\left[y_i\right] = (1 + c|J|)\,\frac{b}{a-1} = 2$$

Given our chosen values of $\{a, b\}$, this restriction reduces to $c = 1/|J|$. Figure 4 plots both sides of (1) for different values of $|J|$ and n. As expected, the asymptotic approximation works very well for high n, for any $|J|$. It also illustrates that in small samples (low n), it's still a pretty good approximation for low $|J|$. It worsens only as there are too many parameters relative to observations, as would be expected from a frequentist perspective. Figure 5 shows that these takeaways apply in log-scale, and for higher sample size.

Figure 4: Expected Variance, Approximation and True Values

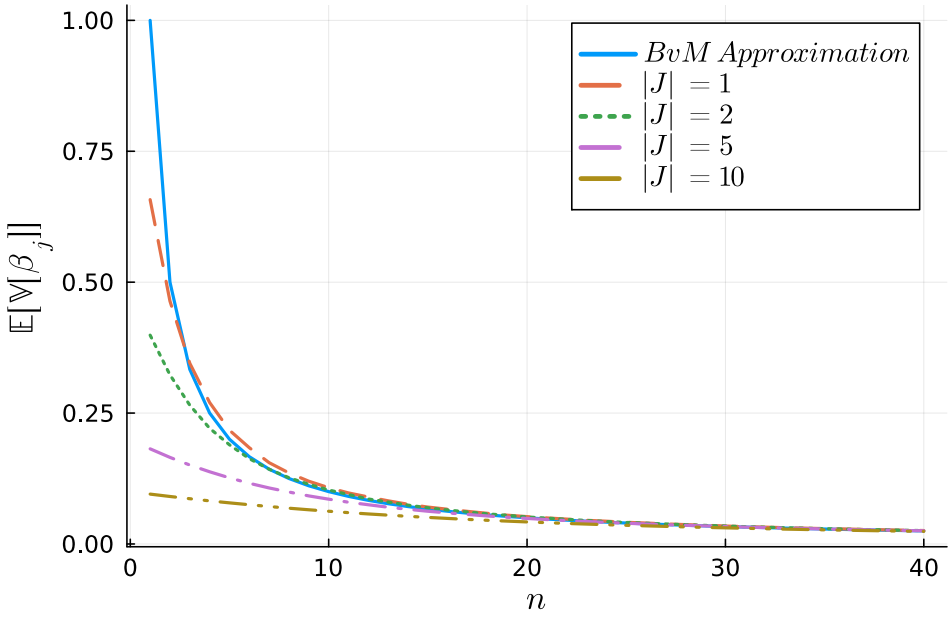
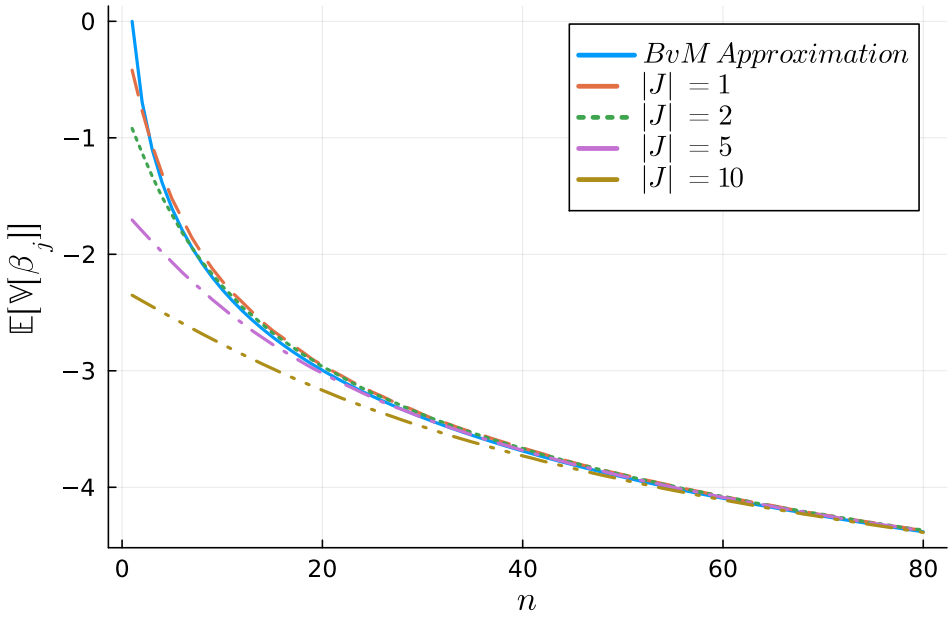


Figure 5 already suggests that for the accuracy of the approximation, all that matters is the proportion of $|J|$ relative to n. This is confirmed by Figures 6 and 7, which show that as we increase $|J|$ in orders of magnitude, if we increase n in the same order (see the x-axes) then the distance between the approximation and the true value of the expected variance seems identical.[14]

[13]The previous statement is true as long as a is sufficiently higher than 1, for example no less than 1.5.

[14]In subplots (b) we start from $n = 10$ and in subplot (c) we start from $n = 100$, to preserve the orders of magnitude

Figure 5: Expected Variance, Approximation and True Values, log-scale



B.2 Bayesian Ex-Ante

Given some $\{k, J, n, \pi\}$, this section shows how to first compute the posterior numerically after observing data $D_J = \{X_J, Y\}$, and then computing the prior loss.

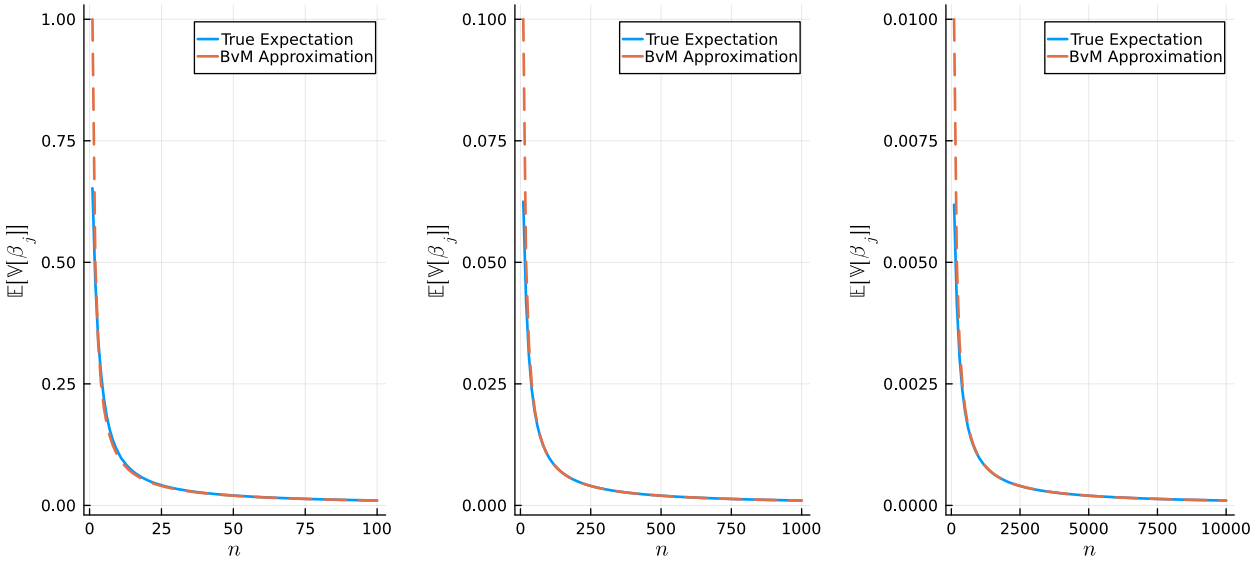
We start by assuming that the prior over β is multivariate normal, where all elements are exchangeable. That is, under π we have $\beta \sim \mathcal{N}_k(\mu, \Sigma)$; where $\mu = c\mathbf{1}$ with $c \in \mathbb{R}$ and $\mathbf{1}$ is a (column) vector of ones of length k; and $\Sigma = \sigma^2 [(1-\rho) I + \rho \mathbf{1}\mathbf{1}']$ where $\sigma^2 > 0$, $\rho \in (0, 1)$, I denotes the identity matrix and $\mathbf{1}'$ denotes the transpose of $\mathbf{1}$. A special case of this formulation is one where $\rho = 0$, in which case all parameters are independent under the prior.

The exchangeability assumption is restrictive, but it allows us to make the simplification of only focusing on the number of covariates in the model $|J|$ as opposed to the specific set of covariates J. It's also the case in which we are most interested in, because the slope of the value of an extra covariate (convexity in the value of information) is meaningful in this case.

To compute a measure of the expected posterior variance over the realizations of data predicted by the prior (the computation of $\mathbb{E}_{Q_\pi}[\cdot]$) we resort to Monte Carlo methods. The algorithm follows:

1. Draw a vector β from $\mathcal{N}_k(\mu, \Sigma)$.
2. Simulate $\{x_i\}_{i=0}^{n-1}$ with n draws from $\mathcal{N}_k(0, I)$.
3. Compute $\{y_i = x_i \beta\}_{i=0}^{n-1}$.
4. Sample draws of $\beta \mid D_J$ from the numerical posterior following a Gibbs sampling algorithm:

Figure 6: Bernstein-von Mises Approximation and True Value



(a) $|J| = 1$ (b) $|J| = 10$ (c) $|J| = 100$

(a) Start β at the true value from step (1). Then, for draw ℓ:

(b) Sample $\beta_{J,\ell}$ from the conditional posterior $\beta_J \mid \beta_{-J,\ell-1}, D_J$. This distribution can be computed analytically following the approach outlined below, in section B.2.1.

(c) Sample $\beta_{-J,\ell}$ from the conditional posterior $\beta_{-J} \mid \beta_{J,\ell}, D_J$. This has to be done computationally, and we follow a standard Metropolis-Hastings procedure, as in Gelman et al. (2013).

(d) Repeat steps 4b-4c a large number of times.

5. Burn the first part of the sample (we burn the initial 5%).

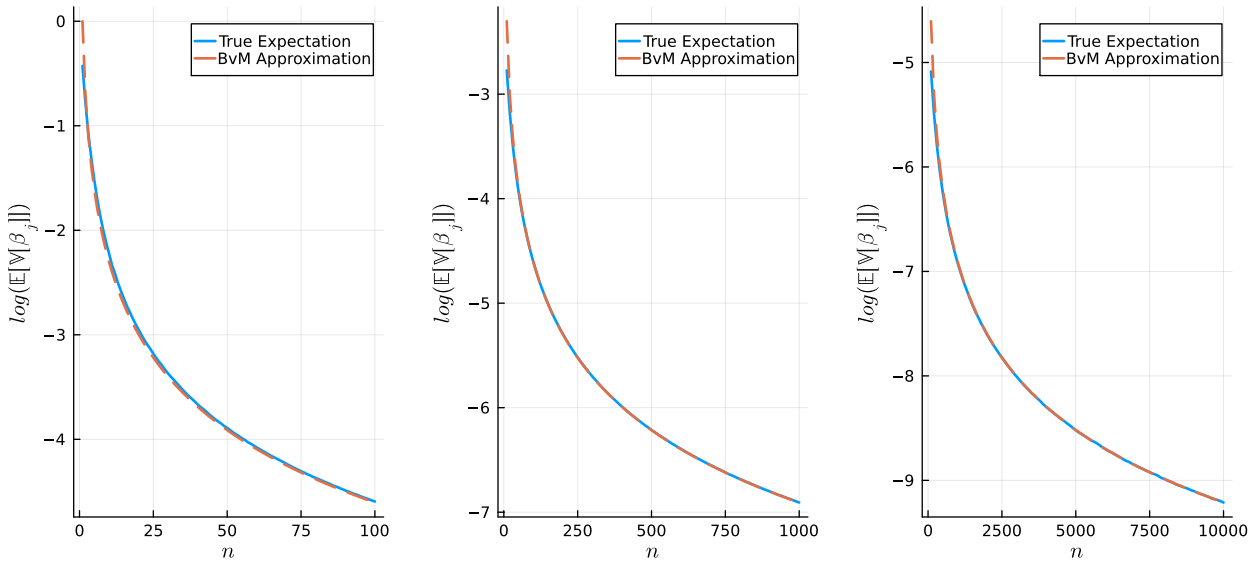
6. Compute the variance of β_J from the posterior samples.

7. Repeat steps 1-6 a large number of times and store results.

8. Compute $\mathbb{E}_{Q_\pi}\left[\mathbb{V}_\pi\left[\beta_j \mid D_J\right]\right]$ by averaging over all $\mathbb{V}_\pi\left[\beta_j \mid D_J\right]$ for $j \in J$.

We then compare the prior loss across agents with different models. Note that in the special case

Figure 7: Bernstein-von Mises Approximation and True Value, log-scale



(a) $|J| = 1$ (b) $|J| = 10$ (c) $|J| = 100$

of $J = \emptyset$, the prior loss can be obtained analytically. From (5):

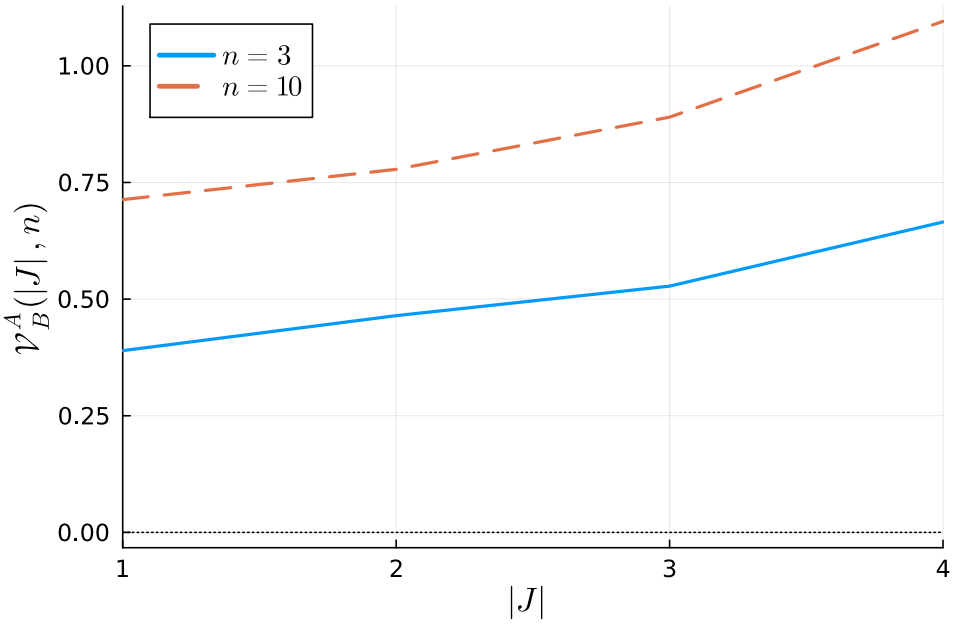
$$\begin{aligned}
\mathcal{L}_B(\emptyset, n) &= \sum_{j=1}^{k} \mathbb{E}_\pi\left[\beta_j^2\right] \\
&= \sum_{j=1}^{k} \left(\mathbb{V}_\pi\left[\beta_j\right] + \mathbb{E}_\pi\left[\beta_j\right]^2\right) \\
&= \sum_{j=1}^{k} \Sigma_{jj}
\end{aligned}$$

No choice of parameters is found to affect our qualitative results, so we illustrate with $k = 5$, $c = 0$, $\rho = 0$ and $\sigma = 1$. Figure 8 shows that even for few observations (and notably, less observations than parameters), the value of an additional covariate is positive and increasing, suggesting a convexity in the value of information.

B.2.1 The conditional distribution of β_J

In order to sample draws from β_J, the Gibbs sampler conditions on β_{-J}. Then, we need to compute the prior for this conditional distribution. This is simple thanks to the functional form of π. Under

Figure 8: Simulated Value of an Additional Covariate



π:

$$\begin{aligned}
\beta_J \mid \beta_{-J} &\sim \mathcal{N}_{|J|}\left(\hat{\mu}_J, \hat{\Sigma}_J\right) \\
\hat{\mu}_J &= \mu_J + \Sigma_{J,-J}\Sigma_{-J,-J}^{-1}\left(\beta_{-J} - \mu_{-J}\right) \\
\hat{\Sigma}_J &= \Sigma_{JJ} - \Sigma_{J,-J}\Sigma_{-J,-J}^{-1}\Sigma_{-J,J}
\end{aligned}$$

Where μ_J follows the same notation at β_J, and Σ_{GH} is the block of Σ corresponding to the set of rows G and the set of columns H. Then, the posterior conditional on the data can be computed analytically as follows:

$$\begin{aligned}
\beta_J \mid \beta_{-J}, D_J &\sim \mathcal{N}_{|J|}\left(\tilde{\mu}_J, \tilde{\Sigma}_J\right) \\
\tilde{\Sigma}_J &= \left(\hat{\Sigma}_J^{-1} + \frac{X_J'X_J}{\sum_{j \notin J}\beta_j^2}\right)^{-1} \\
\tilde{\mu}_J &= \tilde{\Sigma}_J\left(\hat{\Sigma}_J^{-1}\hat{\mu}_J + \frac{X_J'Y}{\sum_{j \notin J}\beta_j^2}\right)
\end{aligned}$$

B.3 Bayesian Ex-Post

This section contains the details regarding the computation of Figure 1. The value of an additional covariate for small samples has a tractable expression from Lemma 2, reproduced here:

$$\mathcal{V}_B\left(D_J, l\right) = \mathbb{E}_\pi\left[\beta_l \mid D_J\right]^2 + \sum_{j \in J'}\left(\mathbb{V}_\pi\left[\beta_j \mid D_J\right] - \mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right] \mid D_J\right]\right) \tag{33}$$

This object can be computed numerically for small n following a small variant of the computational algorithm outlined in appendix B.2. That algorithm allowed us to sample numerically from the posterior distribution, which does not have an analytical form. The only object here that does not follow from that algorithm is $\mathbb{E}_{Q_x}\left[\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right] \mid D_J\right]$ for some $j \in J'$. Given some D_J, this object can be computed by simulating $\{x_{li}\}_{i=0}^{n-1}$ a large number of times, then for each simulation computing $\mathbb{V}_\pi\left[\beta_j \mid D_{J'}\right]$ where D_J is held constant in $D_{J'} = \left\{D_J, \{x_{li}\}_{i=0}^{n-1}\right\}$, and finally averaging over posterior variances.

Then, the computation of $\mathbb{E}_\pi\left[R^2\left(\beta\right) \mid D_J\right]$ follows immediately from sampling from the posterior distribution, following the approach of appendix B.2.

Given a prior π, the full algorithm for this exercise is given by:

1. Sample X from n draws of $\mathcal{N}_k\left(0, I\right)$.
2. Sample some β (see below for details).
3. Compute Y given β.
4. Using the Metropolis-Hasting algorithm with Gibbs Sampling outlined in appendix B.2, sample many draws from the posterior distribution of β given D_J.
5. Compute $\mathbb{E}_\pi\left[R^2\left(\beta\right) \mid D_J\right]$ from the samples drawn in the previous steps.
6. Compute $\mathcal{V}_B\left(D_J, l\right)$ with (33).
7. Repeat steps 1-6 a large number of times.

Our prior of choice puts weight on different values of β while keeping the unconditional variance of y_i constant, in the spirit of corollary 1. We do this by sampling $\left\{\beta_j^2\right\}_{j=1}^k$ from a Dirichlet distribution, with parameters given by a vector of ones of length k. This guarantees that $\sum_{j=1}^k \beta_j^2 = \mathbb{V}\left[y_i\right] = 1$. We choose $k = 3$ for simplicity (so the agent has one covariate and is evaluating whether to purchase one more), and the prior is identical to that in appendix B.2. In particular, we choose $c = 0$, $\rho = 0$ and $\sigma = 1/\sqrt{k}$. The last choice ensures that $\mathbb{V}_\pi\left[y_i\right] = 1$. The results to this exercise are presented in Figure 1.

References

Aina, C. and F. Schneider (2025): "Weighting competing models," Mimeo Universitat Pompeu Fabra.

Alaoui, L. and A. Penta (2016): "Endogenous Depth of Reasoning," *The Review of Economic Studies*, 83, 1297–1333.

Ambuehl, S. and H. C. Thysen (2024): "Choosing between causal interpretations: An experimental study," Mimeo University of Zurich.

Andrei, D., H. Friedman, and N. B. Ozel (2023): "Economic Uncertainty and Investor Attention," *Journal of Financial Economics*, 149, 179–217.

Ba, C. (2025): "Robust Misspecified Models and Paradigm Shifts," arXiv preprint arXiv:2106.12727.

Barbaglia, L., L. Frattarolo, L. Onorante, F. M. Pericoli, M. Ratto, and L. Tiozzo Pezzoli (2023): "Testing Big Data in a Big Crisis: Nowcasting under Covid-19," *International Journal of Forecasting*, 39, 1548–1563.

Barron, K. and T. Fries (2024): "Narrative persuasion," WZB Discussion Paper.

Bloom, N. (2009): "The Impact of Uncertainty Shocks," *Econometrica*, 77, 623–685.

Bohren, J. A. and D. N. Hauser (2024): "Misspecified Models in Learning and Games," *Annual Review of Economics*, 17.

Buell, B., C. Chen, R. Cherif, H. Seo, J. Tang, and N. Wendt (2021): "Impact of COVID-19: Nowcasting and Big Data to Track Economic Activity in Sub-Saharan Africa," SSRN working paper.

Chaloner, K. and I. Verdinelli (1995): "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 273–304.

Cho, I. and K. Kasa (2015): "Learning and Model Validation," *The Review of Economic Studies*, 82, 45–82.

Fischer, P. E., M. S. Heinle, J. M. Kim, and C. Zhu (2024): "Rational Information Acquisition Screens," SSRN working paper.

Fudenberg, D. and G. Lanzani (2023): "Which Misspecifications Persist?" *Theoretical Economics*, 18, 1271–1315.

Gabaix, X. (2014): "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 129, 1661–1710.

——— (2019): "Behavioral Inattention," in *Handbook of Behavioral Economics - Foundations and Applications 2*, ed. by B. D. Bernheim, S. DellaVigna, and D. Laibson, Amsterdam: North-Holland, vol. 2, 261–343.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2023): "Channeled Attention and Stable Errors," Mimeo, Harvard University.

Gelman, A., B. Goodrich, J. Gabry, and A. Vehtari (2019): "R-squared for Bayesian Regression Models," *The American Statistician*, 73, 307–309.

Giacomini, R., V. Skreta, and J. Turen (2020): "Heterogeneity, Inattention, and Bayesian Updates," *American Economic Journal: Macroeconomics*, 12, 282–309.

Hansen, L. P. and T. J. Sargent (2008): *Robustness*, Princeton University Press, student edition ed.

He, K. and J. Libgober (2025): "Misspecified Learning and Evolutionary Stability," *Journal of Economic Theory*, 230, 106082.

Hellwig, C. and L. Veldkamp (2009): "Knowing What Others Know: Coordination Motives in Information Acquisition," *The Review of Economic Studies*, 76, 223–251.

Leamer, E. E. (1978): *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley Series in Probability and Mathematical Statistics, New York: Wiley, a Wiley-Interscience publication; includes index.

Lindley, D. V. (1956): "On a Measure of the Information Provided by an Experiment," *The Annals of Mathematical Statistics*, 27, 986–1005.

Maćkowiak, B., F. Matějka, and M. Wiederholt (2023): "Rational inattention: A review," *Journal of Economic Literature*, 61, 226–273.

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

Myatt, D. P. and C. Wallace (2012): "Endogenous Information Acquisition in Coordination Games," *The Review of Economic Studies*, 79, 340–374.

Ortoleva, P. (2012): "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, 102, 2410–2436.

Sims, C. A. (2003): "Implications of Rational Inattention," *Journal of Monetary Economics*, 50, 665–690.

Veldkamp, L. and C. Chung (2024): "Data and the Aggregate Economy," *Journal of Economic Literature*, 62, 458–484.