The value of conceptual knowledge*

Benjamin Davies† and Anirudh Sankar‡

Draft version: October 26, 2025

Click here for latest version

Abstract

We study the instrumental value of conceptual knowledge when making statistical decisions. Such knowledge tells agents how unknown, payoff-relevant states relate. It is distinct from the statistical knowledge gained from observing signals of those states. We formalize this distinction in a tractable framework used by economists and statisticians. Conceptual knowledge is valuable because it empowers agents to design more informative signals. It is more valuable when states are more "reducible": when they can be explained with fewer common concepts. Its value is non-monotone in the number of signals and vanishes when agents have infinitely many signals. Agents who know more concepts can attain the same payoffs with fewer signals. This is especially true when states are highly reducible.

JEL classification: C44, D83
Keywords: Bayesian learning, concepts, dimension reduction, eigenvalues, models, statistical decisions, value of information

*We thank Steve Callander, Arun Chandrasekhar, Ben Golub, Matt Jackson, Annie Liang, Jann Spiess, and seminar participants at Motu and Stanford for helpful discussions and comments.
†Department of Economics, Stanford University; bldavies@stanford.edu.
‡Department of Economics, Stanford University; asankar@stanford.edu.

1 Introduction

Humans use mental models to make sense of the world (Johnson-Laird, 1983). The building blocks of these models are "concepts": mental representations we use to describe objects in our environment and how they relate (Murphy, 2002). Understanding these relationships empowers us to use information about one object to draw inferences about another (Mitchell, 2021).

This paper studies the interaction between (i) the conceptual knowledge embedded in mental models and (ii) the statistical knowledge inferred from data. We ask: when and why is conceptual knowledge instrumentally "valuable"?

For example, suppose a farmer wants to learn which fertilizers to apply to his crops. He views fertilizers as "black boxes," knowing *that* they help crops grow but not *why*. He does not know how fertilizers' effects relate and cannot extrapolate one from another's. So when he tries different fertilizers to learn their effects, he has to try them separately.

Now suppose the farmer knows fertilizers supply nitrogen, a nutrient that helps crops grow. This tells him how fertilizers' effects relate: they share a common "nitrogen component." He can use this component to extrapolate one fertilizer's effect from another's. Moreover, when he tries different fertilizers to learn their effects, he can combine them to isolate the nitrogen component. This is better than trying each fertilizer separately because it allows him to learn their effects from one trial rather than many.

Nitrogen is a mental construct: the farmer cannot see it. All he can see are the effects of trying different fertilizers. But his conceptual knowledge of nitrogen allows him to learn more efficiently. It empowers him to run trials that are more informative and instrumentally valuable. *How much more valuable* is a quantity we define and characterize in this paper. We call this quantity the "value of conceptual knowledge."

Quantifying this value is important for designing interventions that give people information. In empirical work motivating and guided by this paper (Sankar et al., 2025), we experimentally study the effect of providing farmers conceptual knowledge about fertilizers. We compare farmers given knowledge and data to farmers given data only, and find that farmers given knowledge make more accurate predictions and more profitable decisions. These results provide empirical support for the theoretical predictions made in this paper. The results suggest a way to improve interventions that give people data: we should also give them conceptual tools to interpret data. By explaining how these tools work and quantifying their value, this paper helps in designing interventions that improve people's lives more cost-effectively.

Quantifying the value of conceptual knowledge is also important for comparing human and artificial intelligence (AI). Humans currently have an edge in using concepts to unite seemingly unrelated phenomena: Newton isolated a concept (gravity) that unites apples falling on earth with the orbits of other planets; Watson and Crick isolated a concept (DNA) that unites crime scene investigation with the origins of domesticated rice; Bernoulli isolated a concept (risk aversion) that unites choices in poker with choices between insurance policies. These and other concepts allow humans to learn from limited data (Tenenbaum et al., 2011). In contrast, AI relies on recognizing

patterns in large, rich sets of data (Goodfellow et al., 2016; Halevy et al., 2009). By understanding what conceptual knowledge *is*, and when and why it is valuable, we can better allocate inferential tasks between humans and AI.

Contributions This paper makes three major contributions. First, we formally distinguish two types of knowledge: statistical and conceptual.[1] Statistical knowledge comes from observing patterns in data; conceptual knowledge tells us how data are generated and what patterns to expect *ex ante*. We formalize this distinction in a tractable framework used by other economists and statisticians. This framework gives us a transparent model of how statistical and conceptual knowledge interact. We use this model to derive theoretical results that make testable predictions our empirical work (Sankar et al., 2025) validates.

Second, we use our framework to define and characterize the value of conceptual knowledge. Our definition builds on that for the instrumental value of information (Howard, 1966; Raiffa and Schlaifer, 1961): whereas information is valuable because it leads to better decisions, conceptual knowledge is valuable because it leads to better information. We quantify *how much better* and study this quantity's comparative statics. Our framework admits a closed-form expression for the value of optimally acquired information, allowing us to analyze and clarify how and why this value changes. Moreover, our focus on a specific decision problem allows us to generate insights and intuitions that are not immediate from studying abstract problems (as done by, e.g., Blackwell, 1951, 1953; Frankel and Kamenica, 2019; Whitmeyer, 2024)

Third, we use our framework to formalize what it means to have "deeper" conceptual knowledge. This allows us to compare the marginal values of having deeper knowledge or more data. It also identifies conceptual knowledge as an economic good one may acquire in the same way data are goods one acquires. In this way, we advance the literature on learning and information acquisition that treats conceptual knowledge as fixed and minimally restrictive (e.g., Bardhi, 2024; Callander, 2011; Schwartzstein, 2014).

This paper also contributes to the literatures on model-based learning, model uncertainty and mis-specification, and human and machine reasoning. We explain these contributions after presenting our framework and main results.

Overview Section 2 elaborates on our leading example of a farmer learning about fertilizers.

Section 3 extends the example to a more general setting. We consider a Bayesian agent who makes a statistical decision. His environment contains a collection of unknown, real-valued states. He learns about these states from noisy signals. Then he takes real-valued actions. His loss equals the mean squared difference between the actions and states. He takes the actions that minimize his posterior expected loss.

[1] These types of knowledge relate to the two types of uncertainty identified by Marinacci (2015, p. 1022): "a state uncertainty within models and a model uncertainty across them." We position our paper relative to the model uncertainty literature in Section 8.

The signals give the agent statistical knowledge. His prior on the states encodes his conceptual knowledge. This knowledge tells him how states relate. It allows him to mentally represent states as combinations of "concepts," which we model as eigenvectors of the state vector's prior variance matrix. The corresponding eigenvalues index concepts' explanatory power: an eigenvalue is larger when the corresponding concept explains more of the states' prior variances.

We say states are more "reducible" when their prior variances are explained by fewer common concepts. This happens when the eigenvalues of the prior variance matrix are more spread out. For example, if one eigenvalue is much larger than the others, then the state vector is likely to be close to a one-dimensional subspace of the many-dimensional state space. The more spread out are the eigenvalues, the more the agent can "reduce" the state vector by representing it as a low-dimensional combination of high-dimensional concepts. This dimension reduction is what makes conceptual knowledge valuable.

Section 4 contains preliminary results that we draw upon in later sections. We characterize the instrumental "value" of the agent's signals, derive sharp bounds on this value (see Proposition 1), and formalize what it means for eigenvalues to be "more spread out."

We define and characterize the "value of conceptual knowledge" in Section 5. First, we suppose the agent designs an "optimal sample" containing the most valuable signals. This sample focuses on the concepts with the most explanatory power (see Proposition 2). Next, we consider a counterfactual "naïve" agent who does not know which concepts have more or less explanatory power. This prevents him from focusing on the concepts with the most power. As a result, his optimal sample is less valuable than the conceptually knowledgeable agent's. We derive the values of optimal samples with and without conceptual knowledge, and call their difference the "value" of such knowledge. This difference equals the payoff gain from having conceptual knowledge and using it to design a more valuable sample.

Our first main result (Theorem 1) says that conceptual knowledge is more valuable when states are more reducible. If the states can be explained by a few common concepts, then the agent gains a lot from identifying those concepts and focusing on them when he designs signals (i.e., "asking the right questions"). In contrast, if every concept has the same explanatory power, then the agent gains nothing from identifying those concepts because he designs the same sample that he would if he was naïve.

Our second main result (Theorem 2) says that the value of conceptual knowledge (i) is non-monotone in the number of signals and (ii) vanishes when there are infinitely many signals. If the agent can observe more signals, then he can learn more about the concepts he focuses on, raising the payoff gain from knowing which to focus on. However, having more signals also prompts him to broaden his focus, lowering the gain from knowing which concepts to focus on. The first effect dominates the second when the number of signals is sufficiently small. As it becomes arbitrarily large, the agent's posterior becomes independent of his prior, and so the conceptual knowledge embedded in his prior becomes irrelevant and loses its instrumental value.

In Section 6, we extend our measure of the value of conceptual knowledge to one of "deeper"

knowledge. We suppose the agent knows some, but not all, of the relevant concepts, and refer to the "depth" of his knowledge as the number he knows. Our third main result (Theorem 3) says that deeper conceptual knowledge is weakly more valuable. However, if the agent knows enough concepts, then knowing more yields no additional value because it does not change the optimal sample he designs.

Finally, in Section 7, we study the trade-off between conceptual and statistical knowledge. Our framework gives us a precise language for describing this trade-off: is the agent better off knowing more concepts or having more signals? Our fourth main result (Theorem 4) says that if the agent knows more concepts, then he can attain the same welfare with fewer signals, especially when states are highly reducible. This is because he can design better samples and extract more value from each signal, lowering the number he needs to attain a given welfare target.

Section 8 discusses our modeling assumptions and related literature. Section 9 concludes. Appendix A contains additional discussions and results. Appendix B contains proofs of our mathematical claims.

2 An illustrative example

This section elaborates on the example presented in our introduction. The example is inspired by our empirical work in Uganda, where we study the role that conceptual knowledge plays when farmers learn about fertilizers (Sankar et al., 2025).

Environment A Bayesian farmer wants to learn the effect $\theta_k \in \mathbb{R}$ of applying fertilizer $k \in \{1, 2\}$ to his crops. His prior on $\theta \equiv (\theta_1, \theta_2)$ is a normal distribution with variance $\mathbb{V}(\theta)$. He observes the outcome

$$y = \theta_1 w_1 + \theta_2 w_2 + u$$

of using $w_1 \in \mathbb{R}$ more units of fertilizer 1 and $w_2 \in \mathbb{R}$ more units of fertilizer 2.[2] The vector $w = (w_1, w_2)$ has Euclidean length $\|w\| = 1$ and the error $u \in \mathbb{R}$ is independently normally distributed with variance $\sigma_u^2 > 0$.[3] It captures the randomness in y due to variation in unobserved factors.

The farmer's data $\mathcal{S} \equiv \{(w, y)\}$ comprise the vector w and outcome y. These data are valuable insofar as they make the farmer's beliefs about θ more precise. We measure the value of $\mathcal{S}$ via the mean difference

$$\pi(\mathcal{S}) \equiv \frac{1}{2} \sum_{k=1}^{2} (\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k \mid \mathcal{S}))$$

between the prior and posterior variances of θ_1 and θ_2.

The farmer chooses the vector w that maximizes $\pi(\mathcal{S})$ subject to the constraint $\|w\| = 1$. Intuitively, he chooses the combination of fertilizers that teaches him as much as possible about their

[2] We interpret negative values of w_k as using less of fertilizer k than the farmer uses currently.

[3] We normalize $\|w\| = 1$ so that only the direction of w (and not its magnitude) affects the informativeness of y.

effects. That he can only choose *one* combination reflects the scarcity and cost of relevant data: our Ugandan setting is one of many where humans must learn from limited data.

Conceptual knowledge The farmer knows the two fertilizers supply equal amounts of nitrogen, a nutrient that helps crops grow. He cannot see or touch nitrogen; it is a mental construct. But he can use his conceptual knowledge of nitrogen to express each fertilizer's effect θ_k as the sum of a common "nitrogen effect" and an idiosyncratic effect. He encodes these effects by the scalars

$$\gamma_1 \equiv \frac{\theta_1 + \theta_2}{\sqrt{2}} \quad \text{and} \quad \gamma_2 \equiv \frac{\theta_1 - \theta_2}{\sqrt{2}},$$

allowing him to express the effect vector

$$\theta = \gamma_1 v_1 + \gamma_2 v_2$$

as a linear combination of two unit vectors

$$v_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} \quad \text{and} \quad v_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}.$$

These vectors form an orthonormal basis for the Euclidean space $\mathbb{R}^2$ containing θ. The common and idiosyncratic effects γ_1 and γ_2 are the coordinates of θ over this basis.

The farmer knows v_1 and v_2, but does not know γ_1 or γ_2. Knowing v_1 and v_2 makes learning θ equivalent to learning $\gamma \equiv (\gamma_1, \gamma_2)$. Moreover, since v_1 and v_2 are orthonormal, his data $\mathcal{S}$ have value[4]

$$\pi(\mathcal{S}) = \frac{1}{2}\sum_{k=1}^{2}(\mathbb{V}(\gamma_k) - \mathbb{V}(\gamma_k \mid \mathcal{S})).$$

Thus, knowing about nitrogen allows the farmer to reframe his problem from learning about the effect vector θ to learning about the coordinate vector γ. He does not have to learn each fertilizer's effect separately; instead, he can learn the nitrogen and idiosyncratic effects, and extrapolate the overall effects. This is important because he only has one observation (w, y) from which to infer two unknowns θ_1 and θ_2. Knowing how these unknowns relate (via v_1 and v_2) allows him to learn about both at the same time by choosing w appropriately.

The farmer's choice of w depends on the relative contributions of γ_1 and γ_2 to the prior variances of θ_1 and θ_2. He knows γ_1 contributes more: the fertilizers' effects are mostly determined by how much nitrogen they supply. So he assumes γ_1 and γ_2 are independently distributed with variances $\lambda_1 = \sigma^2(1+\rho)$ and $\lambda_2 = \sigma^2(1-\rho)$. The sum

$$\begin{aligned}\lambda_1 + \lambda_2 &= \mathbb{V}\left(\frac{\theta_1 + \theta_2}{\sqrt{2}}\right) + \mathbb{V}\left(\frac{\theta_1 - \theta_2}{\sqrt{2}}\right)\\ &= \mathbb{V}(\theta_1) + \mathbb{V}(\theta_2)\end{aligned}$$

[4]We derive this expression for $\pi(\mathcal{S})$ in Section 4.1.

of these variances equals the sum of the prior variances of θ_1 and θ_2. The parameter $\rho \in [0, 1)$ determines the share

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1+\rho}{2}$$

of this sum contributed by γ_1. This share equals $1/2$ when $\rho = 0$, in which case γ_1 and γ_2 contribute equally. It equals one in the limit as $\rho \to 1$, in which case only γ_1 contributes. The larger is ρ, the more likely is θ to be close to the one-dimensional subspace of $\mathbb{R}^2$ spanned by v_1.

The coordinate vector $\gamma \equiv (\gamma_1, \gamma_2)$ has variance

$$\mathbb{V}(\gamma) = \sigma^2 \begin{bmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{bmatrix},$$

and so the effect vector

$$\theta = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \gamma$$

has prior variance

$$\begin{aligned} \mathbb{V}(\theta) &= \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right) \mathbb{V}(\gamma) \left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \right)^T \\ &= \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \end{aligned} \tag{1}$$

Thus θ_1 and θ_2 have equal prior variances σ^2 and correlation ρ. Intuitively, the more θ_1 and θ_2 are determined by the common effect γ_1, the more likely they are to have similar values.

The prior variance matrix (1) has eigendecomposition

$$\mathbb{V}(\theta) = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T.$$

Each eigenvalue λ_k equals the prior variance of θ in the direction of the corresponding eigenvector v_k. So θ has the most prior variance in the direction of v_1 and the least in the direction of v_2.

Value of information The value $\pi(\mathcal{S})$ of the farmer's data is largest when $w = \pm v_1$ and smallest $w = \pm v_2$ (see Proposition A2). For example, choosing $w = v_1$ makes $y = \gamma_1 + u$ a "pure signal" of γ_1. This makes $\mathcal{S}$ maximally valuable because it provides information about the component of θ with the most prior variance, leading to the largest difference between prior and posterior variances. In contrast, choosing $w = v_2$ makes $y = \gamma_2 + u$ a pure signal of γ_2. This makes $\mathcal{S}$ *minimally* valuable because it provides information about the component of θ with the *least* prior variance, leading to the *smallest* difference between prior and posterior variances.[5]

[5]In general, the data $\mathcal{S}$ are most valuable when they contain information about the components of θ with the most prior variance. We formalize and prove this claim in Sections 4.1–4.3.

Value of conceptual knowledge The data $\mathcal{S}$ have maximal value

$$\pi^* \equiv \max_{\|w\|=1} \pi(\mathcal{S}).$$

The farmer attains π^* by choosing $w = \pm v_1$. This choice relies on his conceptual knowledge: he must know θ_1 and θ_2 share a common nitrogen effect. Without this knowledge, the farmer would have no way to represent θ_k as the sum of components with differential contributions to its prior variance. So he would assume equal contributions (i.e., $\rho = 0$) and his data would have maximal value

$$\pi^{(0)} \equiv \max_{\|w\|=1} \Big[\pi(\mathcal{S})\big|_{\rho=0}\Big].$$

The difference

$$\Pi \equiv \pi^* - \pi^{(0)}$$

between π^* and $\pi^{(0)}$ captures the value of the farmer's conceptual knowledge: the value of knowing about nitrogen and using this knowledge to make his data more valuable.

The value Π of the farmer's conceptual knowledge is larger when ρ is larger.[6,7] Intuitively, if most of fertilizers' effects come from supplying nitrogen, then the farmer can refine his prior a lot by isolating the nitrogen effect when he tries fertilizers. We formalize this intuition in Section 5, and generalize it to a setting in which $\mathcal{S}$ has arbitrary size and θ has arbitrary length. In this setting, conceptual knowledge is more valuable when the eigenvalues of the prior variance matrix $\mathbb{V}(\theta)$ are more spread out (see Theorem 1). This is why raising ρ raises Π: it raises $\lambda_1 = (1+\rho)\sigma^2$ and lowers $\lambda_2 = (1-\rho)\sigma^2$ without changing their mean $(\lambda_1 + \lambda_2)/2 = \sigma^2$.

3 Framework

We consider a Bayesian agent who collects data before making a statistical decision. This section describes the agent's environment, formalizes his conceptual knowledge about that environment, and presents some specific examples.

3.1 Environment

Prior There is a true but unknown vector $\theta \equiv (\theta_1, \ldots, \theta_K)$ of real-valued "states." The agent's prior on θ is a probability distribution $\mathbb{P}$ over the K-dimensional Euclidean space $\mathbb{R}^K$. This distribution is normal with mean $\mu \in \mathbb{R}^K$ and variance $\Sigma \in \mathbb{R}^{K\times K}$:

$$\mathbb{P} = \mathcal{N}(\mu, \Sigma).$$

[6] We have

$$\pi^* = \frac{(1+\rho)^2\sigma^4}{2\big((1+\rho)\sigma^2 + \sigma_u^2\big)} \quad \text{and} \quad \pi^{(0)} = \frac{\sigma^4}{2(\sigma^2+\sigma_u^2)}$$

by Proposition A2 and the definition of $\pi^{(0)}$. So $\partial\pi^*/\partial\rho > 0$ and $\partial\pi^{(0)}/\partial\rho = 0$, from which it follows that $\partial\Pi/\partial\rho > 0$.

[7] For example, the correlation ρ will be close to one when the fertilizers supply nitrogen only, and close to zero when their nutrient profiles are very different.

We assume $K \geq 2$ is finite and Σ is invertible.

The agent deduces $\mathbb{P}$ from his conceptual knowledge about θ. We explain this deduction in Section 3.2.

Data The agent observes a sample $\mathcal{S} \equiv \{(w^{(i)}, y^{(i)})\}_{i=1}^{n}$ of size n. Each "observation" $(w^{(i)}, y^{(i)})$ comprises a "covariate" $w^{(i)} \in \mathbb{R}^K$ with Euclidean length $\|w^{(i)}\| = 1$,[8] and an "outcome"

$$y^{(i)} = \theta^T w^{(i)} + u^{(i)} \tag{2}$$

equal to the sum of

$$\theta^T w^{(i)} = \sum_{k=1}^{K} \theta_k w_k^{(i)}$$

and an independently normally distributed error $u^{(i)}$ with mean zero and variance $\sigma_u^2 > 0$. Thus, the outcome $y^{(i)}$ provides a noisy signal of a weighted combination of states, where the weights are determined by the covariate $w^{(i)} \equiv (w_1^{(i)}, \ldots, w_K^{(i)})$.

Actions and losses The agent uses his prior $\mathbb{P}$, the sample $\mathcal{S}$, and Bayes' rule to form posterior beliefs about θ. Then he chooses a K-vector $a \equiv (a_1, \ldots, a_K)$ of real-valued actions. These actions induce a loss

$$L(\theta, a) \equiv \frac{1}{K} \sum_{k=1}^{K} (a_k - \theta_k)^2$$

equal to the mean squared difference between them and the corresponding states.[9]

Let $\mathbb{E}$ take expectations with respect to the prior distribution $\mathbb{P}$. The agent chooses the action vector that minimizes his posterior expected loss:[10]

$$a \in \underset{a' \in \mathbb{R}^K}{\arg\min}\, \mathbb{E}[L(\theta, a') \mid \mathcal{S}]. \tag{3}$$

[8] Assuming the covariates have unit length normalizes the scales of the signals $y^{(1)}, \ldots, y^{(n)}$ so that only the directions of $w^{(1)}, \ldots, w^{(n)}$ (and not their magnitudes) affect signals' informativeness. It also ensures the Gram matrix (14) always has trace n (see Footnote 25). This allows us to identify optimal samples of size n via a specific Gram matrix (22).

[9] Suppose $p_1, \ldots, p_K$ are strictly positive and sum to one. Let D be the $K \times K$ diagonal matrix with kk^{th} entry Kp_k, and let $a' \equiv Da$ and $\theta' \equiv D\theta$. Then

$$L(a', \theta') = \sum_{k=1}^{K} p_k (a_k - \theta_k)^2$$

is a weighted average of the squared differences between the actions and corresponding states. The weights $p_1, \ldots, p_K$ encode the agent's preferences: the larger is p_k, the larger is the loss from taking an action a_k different than the state θ_k. We focus on the case with $p_k = 1/K$ for each k, which makes D equal the identity matrix and θ' equal θ. However, we can easily generalize our analysis to a setting with non-equal weights by replacing θ with θ'. Then what matters are the eigenvalues and eigenvectors of $\mathbb{V}(\theta') = D\Sigma D^T$, rather than those of Σ. This does not change our results or insights substantively.

[10] In Appendix Section A1, we explain how the choice problem (3) is equivalent to a prediction problem that arises in the machine and statistical learning literatures. This equivalence comes from interpreting $\theta_1, \ldots, \theta_K$ as values of an unknown function.

Intuitively, he wants to estimate $\theta_1, \ldots, \theta_K$ accurately, and the accuracy of his estimates $a_1, \ldots, a_K$ is determined by their squared errors $(\theta_k - a_k)^2$. Thus, our framework aligns with least squares estimation, a tool used throughout empirical economics and statistics.[11]

Value of $\mathcal{S}$ If the agent did not observe the sample $\mathcal{S}$, then his minimized prior and posterior expected losses would be equal. The information in $\mathcal{S}$ is instrumentally valuable because it helps the agent take actions with lower expected losses. Accordingly, we define the "value of $\mathcal{S}$" to be the difference between his minimized prior and posterior expected losses:[12]

$$\pi(\mathcal{S}) \equiv \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a')] - \min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}]. \tag{4}$$

Covariate selection The agent chooses the covariates $w^{(1)}, \ldots, w^{(n)}$ that maximize (4) subject to the length constraints $\|w^{(i)}\| = 1$. Intuitively, he wants to design the signals $y^{(1)}, \ldots, y^{(n)}$ so that they provide as much payoff-relevant information as possible about the state vector θ.

Timing First, nature draws θ from the prior distribution $\mathbb{P}$. Second, the agent chooses the covariates $w^{(1)}, \ldots, w^{(n)}$ and observes the outcomes $y^{(1)}, \ldots, y^{(n)}$. Third, he combines $\mathbb{P}$ and the sample $\mathcal{S} \equiv \{(w^{(i)}, y^{(i)})\}_{i=1}^n$ to form posterior beliefs about θ. Finally, he chooses the action vector (3) that minimizes his posterior expected loss.

3.2 Conceptual knowledge

Mental model and concepts Our definition of conceptual knowledge draws upon psychologists' and cognitive scientists': concepts are the building blocks of mental models (Johnson-Laird, 1983), are used to describe objects and how they relate (Murphy, 2002), and allow humans to generalize across objects (Mitchell, 2021).

Accordingly, our agent's conceptual knowledge allows him to describe the states $\theta_1, \ldots, \theta_K$ and how they relate. It gives him a mental model of

$$\theta = \sum_{k=1}^{K} \gamma_k v_k \tag{5}$$

as an unknown combination of known vectors $v_1, \ldots, v_K \in \mathbb{R}^K$. We call these vectors "concepts."[13] They are the building blocks of the agent's mental model. They capture his environment's generalizable structure: each state θ_j depends on the j^{th} component of v_k via a coefficient $\gamma_k \in \mathbb{R}$

[11] Indeed, if the prior distribution $\mathbb{P}$ is diffuse, then (3) equals the Ordinary Least Squares estimate of θ given $\mathcal{S}$.

[12] Raiffa and Schlaifer (1961, p. 90) define a similar object and call it the "(expected) value of sample information."

[13] The vectors $v_1, \ldots, v_K$ may not correspond to physical features of the agent's environment. Instead they are mental constructs he uses to make sense of his environment. For example, nutrients like nitrogen are mental constructs: no farmer sees them. All farmers see are the effects of applying fertilizers. This is why we call $v_1, \ldots, v_K$ "concepts."

that is independent of j.[14] This allows the agent to generalize across states: signals of θ_1 provide information about $\gamma_1, \ldots, \gamma_K$, from which he can extrapolate $\theta_2, \ldots, \theta_K$.

For example, suppose $\theta_1, \ldots, \theta_K$ are the effects of applying different fertilizers. Then $v_1, \ldots, v_K$ could encode nutrient quantities and $\gamma_1, \ldots, \gamma_K$ the fertilizer-invariant effects of supplying different nutrients.[15] Moreover, if a farmer learns one fertilizer's overall effect, then he can extrapolate the others' via their joint dependence on $\gamma_1, \ldots, \gamma_K$.

For convenience and without loss of generality, we assume $v_1, \ldots, v_K$ are orthonormal for the remainder of the paper.

Eigendecomposition The agent knows the concepts $v_1, \ldots, v_K$ included in his mental model (5). He does not know the coefficients $\gamma_1, \ldots, \gamma_K$, but he knows some contribute more to the states' prior variances than others. Specifically, he knows each coefficient γ_k is independently distributed with variance $\lambda_k > 0$ non-decreasing in k.[16] Then θ has prior variance

$$\begin{aligned} \Sigma &= V \Lambda V^T \\ &= \sum_{k=1}^{K} \lambda_k v_k v_k^T, \end{aligned} \tag{6}$$

where

$$\Lambda \equiv \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_K \end{bmatrix}$$

is the $K \times K$ diagonal matrix with entries $\lambda_1 \geq \cdots \geq \lambda_K \geq 0$ and

$$V \equiv \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix}$$

is the $K \times K$ orthogonal matrix with columns $v_1, \ldots, v_K$.

Equation (6) is an eigendecomposition of Σ. The k^{th} largest eigenvalue $\lambda_k = \mathbb{V}(\gamma_k)$ of Σ equals the prior variance of θ in the direction of the corresponding unit eigenvector v_k. The trace

$$\text{tr}(\Sigma) = \sum_{k=1}^{K} \lambda_k$$

[14] The coefficients $\gamma_1, \ldots, \gamma_K$ are akin to "deep parameters" that determine the "reduced-form" states $\theta_1, \ldots, \theta_K$ via the structural relationships encoded by $v_1, \ldots, v_K$ (Lucas, 1976).

[15] Alternatively, if $\theta_1, \ldots, \theta_K$ are the prices of financial assets, then $v_1, \ldots, v_K$ could encode payoffs in different states of nature and $\gamma_1, \ldots, \gamma_K$ the prices of Arrow-Debreu securities (see, e.g., Varian, 1987).

[16] It is without loss of generality to assume $\gamma_1, \ldots, \gamma_K$ are independently distributed. This is because Λ is positive-semidefinite, and so, by the spectral theorem, there is an orthogonal matrix $A \in \mathbb{R}^{K \times K}$ and diagonal matrix $\Lambda' \in \mathbb{R}^{K \times K}$ such that $\Lambda = A \Lambda' A^T$. Then $V' \equiv VA$ is orthogonal and Σ has eigendecomposition $V' \Lambda' (V')^T$, so we can carry out our analysis by replacing V with V' and Λ with Λ'. Likewise, it is without loss to assume $\lambda_1 \geq \cdots \geq \lambda_K$ because we can permute the indices of the eigenpairs (λ_k, v_k) without changing Σ.

of Σ equals the sum of the eigenvalues $\lambda_1, \ldots, \lambda_K$. So these eigenvalues' mean

$$\begin{aligned}\overline{\lambda} &\equiv \frac{1}{K}\sum_{k=1}^{K}\lambda_k \\ &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{V}(\theta_k)\end{aligned}$$

equals the mean of the states' prior variances. The ratio $\lambda_k/\operatorname{tr}(\Sigma)$ equals the share of these variances contributed by γ_k. If the shares contributed by $\gamma_1, \ldots, \gamma_K$ are equal, then $\lambda_k = \operatorname{tr}(\Sigma)/K = \overline{\lambda}$ is constant in k and so $\Sigma = V\Lambda V^T$ is proportional to $K \times K$ identity matrix I_K:

$$V\left(\overline{\lambda} I_K\right)V^T = \overline{\lambda} I_K.$$

In contrast, if $\lambda_1/\operatorname{tr}(\Sigma) \approx 1$, then γ_1 contributes most of the states' prior variances.

Reducibility The distribution of $\lambda_1, \ldots, \lambda_K$ around their mean $\overline{\lambda} = \operatorname{tr}(\Sigma)/K$ captures the states' "reducibility." They are more "reducible" when their prior variances are explained by fewer common concepts: when $\lambda_1, \ldots, \lambda_K$ are more spread out around $\overline{\lambda}$.[17,18] The agent's conceptual knowledge allows him to "reduce" the state vector θ by representing it as a low-dimensional combination of higher-dimensional concepts.

If the agent had no conceptual knowledge—i.e., if he did not have a mental model of θ as a combination of concepts with different explanatory powers—then he would not be able to reduce states in the manner described above. His prior variance matrix

$$\Sigma^{(0)} \equiv \overline{\lambda} I_K$$

would equal the prior variance matrix in the case when $\lambda_k = \overline{\lambda}$ for each $k \in \{1, \ldots, K\}$. We refer to $\mathbb{P}^{(0)} \equiv \mathcal{N}(\mu, \Sigma^{(0)})$ as a "naïve" prior because it ignores the covariances among states stemming from their dependence on common concepts.[19]

We can use the true prior $\mathbb{P} \equiv \mathcal{N}(\mu, \Sigma)$ and naïve prior $\mathbb{P}^{(0)}$ to measure how much the agent's conceptual knowledge allows him to reduce states. Since $\mathbb{P}$ and $\mathbb{P}^{(0)}$ are normal distributions with

[17] We formalize what it means for $\lambda_1, \ldots, \lambda_K$ to be "more spread out" in Section 4.4.

[18] If $\lambda_2 = \cdots = \lambda_K$ (as in Example 1), then the distribution of $\lambda_1, \ldots, \lambda_K$ is fully determined by the leading eigenvalue λ_1 and the "spectral gap" $(\lambda_1 - \lambda_2)$. This gap appears elsewhere in the statistical literature: it determines Markov chains' mixing times (Levin et al., 2008) and whether principal components can be estimated consistently (Yu et al., 2015).

[19] The naïve prior variance matrix $\Sigma^{(0)}$ is robust to mis-specification in that it commits as little as possible to any given covariance structure. By spreading variance evenly across all dimensions of $\mathbb{R}^K$, it avoids overweighting components of θ that later prove irrelevant. However, it also forfeits any potential gains from overweighting components that later prove essential.

equal means, the Kullback-Leibler (hereafter "KL") divergence from $\mathbb{P}$ and $\mathbb{P}^{(0)}$ equals[20]

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) &= \frac{1}{2}\left(\mathrm{tr}((\Sigma^{(0)})^{-1}\Sigma) - K + \ln\left(\frac{\det(\Sigma^{(0)})}{\det(\Sigma)}\right)\right) \\ &= -\frac{1}{2}\sum_{k=1}^{K}\ln\left(\frac{\lambda_k}{\overline{\lambda}}\right). \end{aligned} \tag{7}$$

The KL divergence (7) measures the information gain from using $\mathbb{P}$ as a prior rather than $\mathbb{P}^{(0)}$. This information is purely conceptual: it does not depend on the sample $\mathcal{S}$. It comes from knowing how to represent states as low-dimensional combinations of high-dimensional concepts. We study the value of this dimension reduction in Section 5.

The KL divergence (7) equals zero when the eigenvalues $\lambda_1, \ldots, \lambda_K$ of Σ are equal and is larger when they are more spread out (see Proposition A4).[21] This is because spreading $\lambda_1, \ldots, \lambda_K$ narrows the distribution $\mathbb{P}$ to a lower-dimensional subspace of $\mathbb{R}^K$.

For example, consider the prior variance matrix (1) derived in Section 2. This matrix has eigenvalues $\lambda_1 = \sigma^2(1+\rho)$ and $\lambda_2 = \sigma^2(1-\rho)$, which have mean $\overline{\lambda} = \sigma^2$ and become more spread out as $\rho \in [0, 1)$ grows. The KL divergence

$$\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = -\frac{1}{2}\ln(1-\rho^2)$$

from $\mathbb{P}$ to $\mathbb{P}^{(0)}$ equals zero when $\rho = 0$ and grows as ρ grows. If $\rho = 0$, then the true prior $\mathbb{P}$ has equal variance in all directions of $\mathbb{R}^2$, and so the agent gains nothing from knowing the eigenvectors v_1 and v_2 of (1). The larger is ρ, the more concentrated is $\mathbb{P}$ around the subspace spanned by v_1, and so the more the agent gains from knowing v_1 and v_2.

Relationship to PCA The example above illustrates the connection between our ideas and principal component analysis (hereafter "PCA"). PCA is a dimension reduction technique that projects a distribution onto its highest variance dimensions. Traditional PCA estimates these dimensions from data. In contrast, our agent derives them from his conceptual knowledge: he knows which dimensions have the highest variance *before* observing any data. This "pre-data PCA" allows him to collect more instrumentally valuable data; we call this benefit the "value of conceptual knowledge" and quantify it in Sections 5–7.

[20] See Rasmussen and Williams (2006, Section A.5) for a derivation of (7).

[21] If $\mathbb{P}^{(0)}$ has mean $\mu^{(0)} \in \mathbb{R}^K$, then (7) becomes

$$\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = \frac{1}{2}\left(\frac{1}{\overline{\lambda}}\|\mu - \mu^{(0)}\| - \sum_{k=1}^{K}\ln\left(\frac{\lambda_k}{\overline{\lambda}}\right)\right).$$

So even if $\mu^{(0)} \neq \mu$, the KL divergence from $\mathbb{P}$ to $\mathbb{P}^{(0)}$ is non-negative and does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS (see Proposition A4). But it is strictly larger than zero when $\mu^{(0)} \neq \mu$, even if $\lambda_1 = \cdots = \lambda_K$.

3.3 Examples

Below are two examples of how the prior variance matrix Σ encodes conceptual knowledge about the states. The first example generalizes the setting described in Section 2. It builds Σ from first principles, starting with the eigenvalues and eigenvectors. The second example builds Σ from knowledge of how the states are generated, then derives the eigenvalues and eigenvectors.

Example 1 (Pairwise correlated states)**.** Suppose the agent knows each state θ_k has two components: a common component that is proportional to the states' mean and an idiosyncratic component that is independent across states. He encodes the common component by the unit vector

$$v_1 = \frac{1}{\sqrt{K}}\mathbf{1}_K,$$

where $\mathbf{1}_K \equiv (1, \ldots, 1)$ is the K-vector of ones. He encodes the idiosyncratic components by unit vectors $v_2, \ldots, v_K$ that are orthogonal to v_1 and each other. The k^{th} coefficient γ_k in (5) has prior variance

$$\lambda_k = \sigma^2 \begin{cases} 1 + \rho(K-1) & \text{if } k = 1 \\ 1 - \rho & \text{if } k > 1, \end{cases}$$

where $\sigma^2 > 0$ is the mean of $\lambda_1, \ldots, \lambda_K$ and where $\rho \in [0, 1)$ determines the share

$$\frac{\lambda_1}{\lambda_1 + \cdots + \lambda_K} = \frac{1}{K} + \rho\left(1 - \frac{1}{K}\right)$$

of the prior variances of $\theta_1, \ldots, \theta_K$ contributed by the coefficient γ_1 on v_1. This share equals $1/K$ when $\rho = 0$, in which case λ_k is constant in k and so $\gamma_1, \ldots, \gamma_K$ contribute to the prior variances of $\theta_1, \ldots, \theta_K$ equally. It equals one in the limit as $\rho \to 1$, in which case only γ_1 contributes.

Since $v_1, \ldots, v_K$ are orthonormal, the sum

$$\sum_{k=1}^{K} v_k v_k^T = I_K$$

of their outer products equals the $K \times K$ identity matrix. Therefore, the prior variance matrix

$$\begin{aligned} \Sigma &= \lambda_1 v_1 v_1^T + \lambda_K\left(I_K - v_1 v_1^T\right) \\ &= \rho\sigma^2 \mathbf{1}_K \mathbf{1}_K^T + (1-\rho)\sigma^2 I_K \\ &= \sigma^2 \begin{bmatrix} 1 & \rho & \cdots \\ \rho & 1 & \\ \vdots & & \ddots \end{bmatrix} \end{aligned} \tag{8}$$

is the $K \times K$ matrix with diagonal entries equal to σ^2 and off-diagonal entries equal to $\rho\sigma^2$. Thus, under the agent's prior, the states have equal variances σ^2 and pairwise correlations ρ.

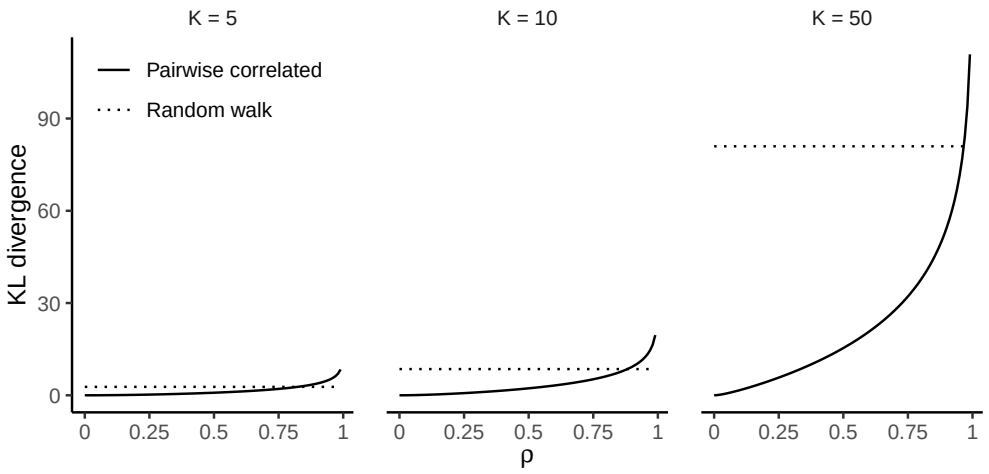

Figure 1: KL divergences $\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)})$ when states are pairwise correlated (i.e., when θ has prior variance (8)) and when they follow a random walk (i.e., when θ has prior variance (9) with $\nu^2 = 2\sigma^2/(K+1)$)

Example 2 (Random walk)**.** Let $\nu > 0$. Suppose the agent knows $\theta_1, \ldots, \theta_K$ are values of a random walk with known initial value $\theta_0 \in \mathbb{R}$ and unknown, independently distributed increments

$$\theta_k - \theta_{k-1} \sim \mathcal{N}(0, \nu^2).$$

Then the prior variance matrix

$$\Sigma = \nu^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & & \vdots \\ 1 & 2 & \cdots & K \end{bmatrix} \tag{9}$$

has jk^{th} entry $\Sigma_{jk} = \nu^2 \min\{j, k\}$. Fortiana and Cuadras (1997) show that (9) has k^{th} largest eigenvalue

$$\lambda_k = \frac{\nu^2}{4} \csc^2\left(\frac{(2k-1)\pi}{2K+1}\right)$$

and that the corresponding unit eigenvector v_k has j^{th} component

$$[v_k]_j = \frac{2}{\sqrt{2K+1}} \sin\left(\frac{j(2k-1)\pi}{2K+1}\right).$$

The eigenvalues of (8) have mean σ^2, whereas the eigenvalues of (9) have mean $\nu^2(K+1)/2$. Choosing $\nu^2 = 2\sigma^2/(K+1)$ equates these two means but does not equate the eigenvalues' distributions, nor the KL divergences $\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)})$ those distributions imply. We illustrate this fact in Figure 1. It shows that assuming states follow a random walk is equivalent, in terms of how much

prior structure it imposes, to assuming a large pairwise correlation.[22] This is especially true when there are many states: if $K = 5$, then the equivalent correlation is about 0.82; if $K = 50$, then it is about 0.97.

4 Preliminaries

This section contains preliminary results that we draw upon in later sections. We characterize the optimal action vector (3), the posterior expected loss it induces, and the value (4) of the sample $\mathcal{S}$. We establish sharp lower and upper bounds on this value, and explain how the agent constructs "optimal" samples. Finally, we formalize what it means for the eigenvalues of the prior variance matrix to be "more spread out."

4.1 Value of $\mathcal{S}$

Let $\mathbb{V}$ take variances with respect to the prior distribution $\mathbb{P}$. Lemma 1 characterizes the optimal action vector (3) and the posterior expected loss it induces.[23] This vector equals the posterior mean of θ. It induces a posterior expected loss equal to the mean of the posterior variances of $\theta_1, \ldots, \theta_K$.

Lemma 1. *The optimal action vector $a = \mathbb{E}[\theta \mid \mathcal{S}]$ induces posterior expected loss*

$$\mathbb{E}[L(\theta, a) \mid \mathcal{S}] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{V}(\theta_k \mid \mathcal{S}). \tag{10}$$

If the sample $\mathcal{S}$ was empty, then (10) would equal the minimized prior expected loss

$$\min_{a' \in K} \mathbb{E}[L(\theta, a')] = \frac{1}{K} \sum_{k=1}^{K} \mathbb{V}(\theta_k). \tag{11}$$

Substituting (10) and (11) into (4) yields and expression for the sample's value $\pi(\mathcal{S})$ in terms of the states' prior and posterior variances:

$$\pi(\mathcal{S}) = \frac{1}{K} \sum_{k=1}^{K} (\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k \mid \mathcal{S})). \tag{12}$$

[22] Callander (2011) and others use Brownian motions (the continuous-time analogues of random walks) as tools for modeling "complexity." They define "complex" environments as those in which only local learning is possible: learning a state provides some information about nearby states but little about distant states (see also Bardhi (2024)). The limiting case is when learning a state provides *no* information about others; in our framework, this happens when the states are uncorrelated. Yet Figure 1 suggests that Brownian motions are as structurally restrictive as assuming states are highly correlated.

[23] Lemma 1 holds even when $\mathbb{P}$ is not normal. We assume $\mathbb{P}$ is normal so that we can derive closed-form expressions for the posterior variances of $\theta_1, \ldots, \theta_K$ and, thus, the value (12) of $\mathcal{S}$.

Intuitively, the sample is valuable insofar as it lowers states' variances, allowing the agent to estimate them more accurately. Moreover, since $v_1, \ldots, v_K$ are known and orthonormal, we have[24]

$$\pi(\mathcal{S}) = \frac{1}{K} \sum_{k=1}^{K} (\mathbb{V}(\gamma_k) - \mathbb{V}(\gamma_k \mid \mathcal{S})). \tag{13}$$

Thus, equivalently, the sample is valuable insofar as it lowers the variances of $\gamma_1, \ldots, \gamma_K$.

4.2 Bounds on $\pi(\mathcal{S})$

We can express (12) in terms of the traces of the prior and posterior variance matrices:

$$\pi(\mathcal{S}) = \frac{1}{K} (\operatorname{tr}(\Sigma) - \operatorname{tr}(\mathbb{V}(\theta \mid \mathcal{S}))).$$

Lemma 2 characterizes $\mathbb{V}(\theta \mid \mathcal{S})$ in terms of the prior variance matrix Σ and "Gram matrix"

$$G \equiv \sum_{i=1}^{n} w^{(i)} (w^{(i)})^T. \tag{14}$$

Lemma 2. *We have*

$$\mathbb{V}(\theta \mid \mathcal{S}) = \left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G \right)^{-1}. \tag{15}$$

The Gram matrix (14) is symmetric and positive semi-definite. So, by the spectral theorem, there is a $K \times K$ diagonal matrix

$$\Delta \equiv \begin{bmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_K \end{bmatrix}$$

with entries $\delta_1 \geq \cdots \geq \delta_K \geq 0$ and a $K \times K$ orthogonal matrix

$$\Omega = \begin{bmatrix} \omega_1 & \cdots & \omega_K \end{bmatrix}$$

such that

$$\begin{aligned} G &= \Omega \Delta \Omega^T \\ &= \sum_{k=1}^{K} \delta_k \omega_k \omega_k^T. \end{aligned} \tag{16}$$

[24] We have $\theta = V\gamma$, where V is the orthogonal matrix with known columns $v_1, \ldots, v_K$ and γ is the vector of unknown coefficients $\gamma_1, \ldots, \gamma_K$. So the prior variances of $\theta_1, \ldots, \theta_K$ and $\gamma_1, \ldots, \gamma_K$ have equal sums:

$$\sum_{k=1}^{K} \mathbb{V}(\theta_k) = \operatorname{tr}(\mathbb{V}(\theta)) = \operatorname{tr}(\mathbb{V}(V\gamma)) = \operatorname{tr}\Big(V\, \mathbb{V}(\gamma) V^T\Big) \overset{\star}{=} \operatorname{tr}(\mathbb{V}(\gamma)) = \sum_{k=1}^{K} \mathbb{V}(\gamma_k),$$

where $\star$ uses the cyclic property of matrix traces and the orthogonality of V. Similarly, the posterior variances of $\theta_1, \ldots, \theta_K$ and $\gamma_1, \ldots, \gamma_K$ have equal sums. Substituting these sums into (12) yields (13).

Then $\delta_1, \ldots, \delta_K$ are the eigenvalues of G and $\omega_1, \ldots, \omega_K \in \mathbb{R}^K$ are the corresponding unit eigenvectors. Proposition 1 uses the eigendecompositions (6) and (16) of the prior variance and Gram matrices to provide sharp bounds on $\pi(\mathcal{S})$.

Proposition 1. *The value $\pi(\mathcal{S})$ of $\mathcal{S}$ satisfies*

$$\frac{1}{K}\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2}\right)^{-1}\right) \overset{\star}{\leq} \pi(\mathcal{S}) \overset{\star\star}{\leq} \frac{1}{K}\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1}\right), \tag{17}$$

where $\star$ holds with equality if $\omega_k = v_{K-k+1}$ for each $k \in \{1, \ldots, K\}$ and $\star\star$ holds with equality if $\omega_k = v_k$ for each $k \in \{1, \ldots, K\}$.

Proposition 1 says that the sample $\mathcal{S}$ is most valuable when the eigenvectors of Σ and G are maximally "aligned": when $v_k = \omega_k$ for each $k \in \{1, \ldots, K\}$ and hence $V = \Omega$. Then $\mathcal{S}$ contains more information about components of θ with larger prior variances. In contrast, the sample is *least* valuable when the eigenvectors of Σ and G are maximally "*mis*-aligned": when $v_k = \omega_{K-k+1}$ for each $k \in \{1, \ldots, K\}$. Then $\mathcal{S}$ contains *less* information about components of θ with larger prior variances.

4.3 Optimal samples

Suppose the eigenvectors of Σ and G are maximally aligned (and hence $V = \Omega$). Then, by Proposition 1, the value $\pi(\mathcal{S})$ of $\mathcal{S}$ rises when the trace

$$\operatorname{tr}(\mathbb{V}(\theta \mid \mathcal{S})) = \sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1}$$

of the posterior variance matrix falls. This trace depends on the eigenvalues $\delta_1, \ldots, \delta_K$ of G, which are non-negative, non-increasing, and sum to n.[25] So $\pi(\mathcal{S})$ is maximized when $\delta_1, \ldots, \delta_K$ solve

$$\min_{\delta_1, \ldots, \delta_K \in \mathbb{R}} \sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1} \text{ subject to } \delta_1 \geq \ldots \geq \delta_K \geq 0 \text{ and } \sum_{k=1}^{K}\delta_k = n. \tag{18}$$

Proposition 2 describes a solution to (18). It uses the integer

$$R^* \equiv \max\left\{k \in \{1, \ldots, K\} : \sum_{j=1}^{k}\frac{1}{\lambda_j} + \frac{n}{\sigma_u^2} \geq \frac{k}{\lambda_k}\right\} \tag{19}$$

to provide a sharp upper bound

$$\pi^* \equiv \frac{1}{K}\left(\sum_{k=1}^{R^*}\lambda_k - (R^*)^2\left(\sum_{k=1}^{R^*}\frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \tag{20}$$

[25] Indeed

$$\sum_{k=1}^{K}\delta_k = \operatorname{tr}(G) = \operatorname{tr}\left(\sum_{i=1}^{n} w^{(i)}(w^{(1)})^T\right) \overset{\star}{=} \sum_{i=1}^{n}\operatorname{tr}\left((w^{(i)})^T w^{(i)}\right) \overset{\star\star}{=} n,$$

where $\star$ uses the linearity and cyclic property of matrix traces, and $\star\star$ uses the fact that $\|w^{(i)}\| = 1$ for each i.

on the value of $\mathcal{S}$.[26]

Proposition 2. *Define*

$$\delta_k^* \equiv \begin{cases} \frac{n}{R^*} + \sigma_u^2\left(\frac{1}{R^*}\sum_{j=1}^{R^*}\frac{1}{\lambda_j} - \frac{1}{\lambda_k}\right) & \text{if } k \leq R^* \\ 0 & \text{if } k > R^* \end{cases} \tag{21}$$

for each $k \in \{1, \ldots, K\}$. Then $\pi(\mathcal{S}) \leq \pi^*$ with equality if

$$G = \sum_{k=1}^{K} \delta_k^* v_k v_k^T. \tag{22}$$

We call the sample "optimal" if it induces the Gram matrix (22). The agent can construct such a sample as follows: for each $k \in \{1, \ldots, K\}$, collect δ_k^* observations with covariate v_k.[27] Then the outcomes $y^{(1)}, \ldots, y^{(n)}$ are pure signals of the coefficients $\gamma_1, \ldots, \gamma_K$. For example, suppose $w^{(1)} = v_1$. Then, since $v_1, \ldots, v_K$ are orthonormal, we have

$$\begin{aligned} y^{(1)} &= \theta^T w^{(1)} + u^{(1)} \\ &= \left(\sum_{k=1}^{K} \gamma_k v_k\right)^T v_1 + u^{(1)} \\ &= \gamma_1 + u^{(1)} \end{aligned}$$

An optimal sample contains pure signals of $\gamma_1, \ldots, \gamma_{R^*}$, the coefficients in (5) that contribute most to the states' prior variances. However, it provides no information about $\gamma_{R^*+1}, \ldots, \gamma_K$, the coefficients that contribute least to the states' prior variances. Thus, the agent optimally "regularizes" by focusing on the coefficients that "matter" and ignoring those that do not. The number R^* that "matter" grows as the sample size n grows. We call R^* the "rank" of an optimal sample because it is the rank of the Gram matrix (22).

If $\mathcal{S}$ is optimal, then the posterior variance matrix $\mathbb{V}(\theta \mid \mathcal{S})$ has k^{th} largest eigenvalue

$$\left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2}\right)^{-1} = \begin{cases} R^*\left(\sum_{j=1}^{R^*}\frac{1}{\lambda_j} + \frac{n}{\sigma_u^2}\right)^{-1} & \text{if } k \leq R^* \\ \lambda_k & \text{if } k > R^* \end{cases}$$

and trace[28]

$$\sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2}\right)^{-1} = (R^*)^2\left(\sum_{k=1}^{R^*}\frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1} + \sum_{k>R^*}\lambda_k. \tag{23}$$

[26]Proposition 2 echoes Liang et al.'s (2022) Theorem 1, which says that if there are two unknown states (which Liang et al. call "attributes"), then one should prioritize learning about the state with more prior variance.

[27]This may be infeasible for two reasons: (i) the eigenvalues $d_1^*, \ldots, d_K^*$ may not be integers; (ii) the agent may not be able to choose $v_1, \ldots, v_K$ as covariates (since, e.g., it would require him to combine negative quantities of fertilizers). We abstract from these issues for convenience and expositional clarity.

[28]The two terms on the RHS of (23) correspond to the "sampling" and "extrapolation" errors discussed in Appendix Section A2.3.

The eigenvalues of $\mathbb{V}(\theta \mid \mathcal{S})$ are the posterior variances of the unknown coefficients $\gamma_1, \ldots, \gamma_K$. So if $\mathcal{S}$ is optimal, then it equates the posterior variances of $\gamma_1, \ldots, \gamma_{R^*}$ to each other and the posterior variances of $\gamma_{R^*+1}, \ldots, \gamma_K$ to their prior variances.[29] Intuitively, the agent has a target variance and designs $\mathcal{S}$ so as to bring the posterior variances of $\gamma_1, \ldots, \gamma_K$ below that target.[30] This minimizes the trace (23) given the sample size n.

4.4 Mean-preserving spreads

Finally, consider the eigenvalues $\lambda_1, \ldots, \lambda_K$ of the prior variance matrix Σ. Let $F : (0, \infty) \to [0, 1]$ be their (empirical) cumulative distribution function (hereafter "CDF"):

$$F(z) = \frac{|\{k \in \{1, \ldots, K\} : \lambda_k \leq z\}|}{K} \tag{24}$$

for all $z > 0$. A "mean-preserving spread" (hereafter "MPS") of F is a CDF $F' : (0, \infty) \to [0, 1]$ such that

(i) The distributions described by F and F' have the same mean:

$$\int_0^\infty z \, \mathrm{d}F(z) = \int_0^\infty z \, \mathrm{d}F'(z).$$

(ii) For all $z > 0$, the area under F' from 0 to z is at least the area under F from 0 to z:

$$\int_0^z \big(F'(t) - F(t)\big) \, \mathrm{d}t \geq 0.$$

These are the "integral conditions" from Rothschild and Stiglitz (1970). Condition (ii) says that F' has more weight in its tails than F, capturing the idea of eigenvalues being more spread out.

We say $\lambda_1, \ldots, \lambda_K$ "undergo a MPS" when their CDF (24) undergoes a MPS. This changes the trace of the posterior variance matrix without changing the trace of Σ. So if $\lambda_1, \ldots, \lambda_K$ undergo a MPS, then the agent's posterior expected loss changes but his prior expected loss does not. This makes MPSs useful for analyzing how the value of $\mathcal{S}$ depends on the distribution of $\lambda_1, \ldots, \lambda_K$. We discuss this dependence in Section 5 and Appendix Section A2, in which we state results that depend on the following lemma:

Lemma 3. *Let $\lambda_k > 0$ and $\lambda'_k > 0$ be non-increasing in $k \in \{1, \ldots, K\}$, and let F and F' be their CDFs defined as in* (24)*. The following are equivalent:*

(i) F' is a mean-preserving spread of F.

[29] This equality of large eigenvalues and ignorance of small eigenvalues is reminiscent of Arrow's (1963) theorem on the optimality of deductible insurance contracts. Such contracts second-degree stochastically dominate all other contracts with the same premia (Gollier and Schlesinger, 1996). They provide full coverage against risks above a minimum threshold. Similarly, optimal samples provide "full coverage against posterior variance" above a minimum threshold.

[30] This strategy is called "reverse water-filling" in rate-distortion theory—see Cover and Thomas (2006, Chapter 10). It also appears in Ilut and Valchev's (2025) model of abstract reasoning.

(ii) $\sum_{k=1}^{K} g(\lambda'_k) \geq \sum_{k=1}^{K} g(\lambda_k)$ *for all convex functions* $g : (0, \infty) \to \mathbb{R}$.

(iii) $\sum_{j=1}^{k} \lambda'_j \geq \sum_{j=1}^{k} \lambda_j$ *for each* $k \in \{1, \ldots, K\}$, *with equality when* $k = K$.

(iv) $\sum_{j=k}^{K} \lambda'_j \leq \sum_{j=k}^{K} \lambda_j$ *for each* $k \in \{1, \ldots, K\}$, *with equality when* $k = 1$.

For example, consider the prior variance matrix (8) constructed in Example 1. This matrix has eigenvalues $\lambda_1 = (1 + \rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$. Their k^{th} partial sum

$$\sum_{j=1}^{k} \lambda_j = (k + \rho(K-k))\sigma^2$$

is increasing in ρ when $k < K$ and constant in ρ when $k = K$. Thus, by Lemma 3, the eigenvalues of (8) undergo a MPS when ρ rises.

5 Value of conceptual knowledge

Whereas information is valuable insofar as it helps the agent make better decisions (i.e., take actions $a_1, \ldots, a_K$ that estimate the states $\theta_1, \ldots, \theta_K$ more accurately), conceptual knowledge is valuable insofar as it helps him obtain better information.

We formalize this idea as follows. Suppose the agent collects an optimal sample with value π^* (see Section 4.3). Designing this sample relies on his conceptual knowledge: his mental model (5) of θ as a combination of concepts with different explanatory powers. If the agent did not have this knowledge, then he would use the "naïve" prior $\mathbb{P}^{(0)} = \mathcal{N}(\mu, \Sigma^{(0)})$ described in Section 3.2. He would assume θ has prior variance $\Sigma^{(0)} = \overline{\lambda} I_K$, a matrix with eigenvalues $\lambda_1^{(0)} = \cdots = \lambda_K^{(0)} = \overline{\lambda}$. So, by analogy to (19) and (20), the agent's optimal sample would have rank

$$\begin{aligned} R^{(0)} &\equiv \max\left\{ k \in \{1, \ldots, K\} : \sum_{j=1}^{k} \frac{1}{\lambda_j^{(0)}} + \frac{n}{\sigma_u^2} \geq \frac{k}{\lambda_k^{(0)}} \right\} \\ &= K \end{aligned}$$

and value

$$\begin{aligned} \pi^{(0)} &\equiv \frac{1}{K}\left(\sum_{k=1}^{R^{(0)}} \lambda_k^{(0)} - (R^{(0)})^2 \left(\sum_{k=1}^{R^{(0)}} \frac{1}{\lambda_k^{(0)}} + \frac{n}{\sigma_u^2} \right)^{-1} \right) \\ &= \frac{\overline{\lambda}\tau}{K + \tau}, \end{aligned}$$

where

$$\tau \equiv \frac{n/\sigma_u^2}{1/\overline{\lambda}}$$

indexes the precision of the data relative to the agent's prior. This sample would provide equal information about each state: the optimal Gram matrix (22) would have equal eigenvalues.[31] Intuitively, if the agent did not know which concepts had more explanatory power, then he would have no reason to prioritize some components of θ over others when collecting data, and so he would collect the same amount on every component.

The true and naïve prior variance matrices Σ and $\Sigma^{(0)}$ have equal traces, and so replacing the true prior $\mathbb{P} = \mathcal{N}(\mu, \Sigma)$ with the naïve prior $\mathbb{P}^{(0)} = \mathcal{N}(\mu, \Sigma^{(0)})$ does not change the agent's prior expected loss (see Section 4.1). But the two priors imply different optimal samples and different posterior expected losses (10). The difference

$$\begin{aligned}\Pi &\equiv \pi^* - \pi^{(0)} \\ &= \frac{\overline{\lambda}}{K}\left(\sum_{k=1}^{R^*}\frac{\lambda_k}{\overline{\lambda}} - (R^*)^2\left(\sum_{k=1}^{R^*}\frac{\overline{\lambda}}{\lambda_k} + \tau\right)^{-1} - \frac{K\tau}{K+\tau}\right)\end{aligned}$$

between π^* and $\pi^{(0)}$ equals the decline in the agent's posterior expected loss from having conceptual knowledge and using it to design an optimal sample.

Accordingly, we call Π the "value of conceptual knowledge." It depends on the eigenvalues $\lambda_1, \ldots, \lambda_K$ of Σ and the precision parameter τ (which jointly determine R^*). We characterize this dependence in Theorems 1 and 2.

Theorem 1. *The value* Π *of conceptual knowledge*

(i) is non-negative,

(ii) equals zero when $\lambda_1, \ldots, \lambda_K$ *are equal, and*

(iii) does not fall when $\lambda_1, \ldots, \lambda_K$ *undergo a MPS.*

Theorem 1 says that conceptual knowledge is more valuable when states are more reducible. If a few common concepts explain most of states' prior variances, then the agent gains a lot from identifying those concepts and learning about their effects $\gamma_1, \ldots, \gamma_{R^*}$ (i.e., "asking the right questions"). In contrast, if every concept has the same explanatory power, then the agent gains nothing from identifying those concepts because he designs the same sample as he would if he was naïve.

Our proof of Theorem 1 draws upon the following three observations. First, every distribution of $\lambda_1, \ldots, \lambda_K$ is a MPS of the degenerate distribution under which they are equal (to their mean $\overline{\lambda}$). Second, if $\lambda_k = \overline{\lambda}$ for each $k \in \{1, \ldots, K\}$, then the agent's optimal sample has the same rank $R^* = R^{(0)}$ and value $\pi^* = \pi^{(0)}$ as a naïve agent's, which implies $\Pi = 0$. Third, whereas $\pi^{(0)}$ does not change when $\lambda_1, \ldots, \lambda_K$ undergo a MPS, the value π^* of an optimal sample can rise but cannot

[31] Indeed, if $\lambda_1, \ldots, \lambda_K$ are equal (to $\overline{\lambda}$), then $\delta_1^*, \ldots, \delta_K^*$ are also equal (to n/K).

fall (see Lemma B3).[32]

Theorem 2. *There is a finite threshold $\tau' \geq 0$ such that Π is increasing in τ if and only if $\tau < \tau'$. This threshold equals zero if and only if $\lambda_1, \ldots, \lambda_K$ are equal. Moreover,*

$$\lim_{\tau \to \infty} \Pi = 0.$$

Theorem 2 says that the value of conceptual knowledge is non-monotone in τ (and thus $n \equiv \tau \sigma_u^2 / \overline{\lambda}$). This is because raising τ has two effects:[33]

(i) it gives the agent more information about the unknown coefficients $\gamma_1, \ldots, \gamma_{R^*}$, raising the gain from knowing which concepts to focus on;

(ii) it leads the agent to learn about more coefficients (i.e., it raises R^*), lowering the gain from knowing which concepts to focus on.

The first effect dominates the second precisely when $\tau < \tau'$. The threshold τ' equals zero if and only if $\lambda_1, \ldots, \lambda_K$ are equal, in which case conceptual knowledge has no value because it does not change the optimal sample from what a naïve agent would design.

Theorem 2 also says that the value of conceptual knowledge vanishes as τ (and thus n) grows without bound. This is because the agent's posterior becomes less dependent on his prior as τ grows and is independent in the limit as $\tau \to \infty$. Intuitively, if the agent has infinite data, then he does not benefit from doing "pre-data PCA" (see Section 3.2) because he can do traditional (post-data) PCA. Having access to unlimited data washes out the benefit of knowing what data to collect. But this washout relies on having *unrestricted* access: the agent must be able to choose

[32] In contrast, the rank R^* of an optimal sample can rise or fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. For example, suppose $K = 3$ and

$$\lambda_k = \begin{cases} 5 & \text{if } k = 1 \\ 3 + \zeta & \text{if } k = 2 \\ 2 - \zeta & \text{if } k = 3 \end{cases}$$

with $\zeta \in \{0, 1\}$. Then $\lambda_1, \lambda_2, \lambda_3$ undergo a MPS when ζ rises from zero to one. Now if $\zeta = 0$ then

$$R^*\big|_{\zeta=0} = \begin{cases} 1 & \text{if } \frac{n}{\sigma_u^2} < \frac{2}{15} \\ 2 & \text{if } \frac{2}{15} \leq \frac{n}{\sigma_u^2} < \frac{7}{15} \\ 3 & \text{if } \frac{7}{15} \leq \frac{n}{\sigma_u^2}, \end{cases}$$

whereas if $\zeta = 1$ then

$$R^*\big|_{\zeta=1} = \begin{cases} 1 & \text{if } \frac{n}{\sigma_u^2} < \frac{1}{20} \\ 2 & \text{if } \frac{1}{20} \leq \frac{n}{\sigma_u^2} < \frac{31}{20} \\ 3 & \text{if } \frac{31}{20} \leq \frac{n}{\sigma_u^2}. \end{cases}$$

So the MPS raises R^* when $1/20 \leq n/\sigma_u^2 < 2/15$ but lowers R^* when $7/15 \leq n/\sigma_u^2 < 31/20$.

[33] Intuitively, raising τ (i) gives the agent better answers to his chosen questions, but also (ii) prompts him to ask more questions. The first effect raises the value of knowing which questions to ask; the second effect lowers it.

covariates $w^{(1)}, \ldots, w^{(n)}$ that span the K-dimensional Euclidean space containing the state vector θ. If the covariates do not span $\mathbb{R}^K$, then $\mathcal{S}$ may contain no information about some components of θ, the agent's posterior expected loss may be arbitrarily large, and the value of $\mathcal{S}$ may be arbitrarily small. We illustrate this possibility in Appendix Section A2.3.

As an illustration of Theorems 1 and 2, consider the prior variance matrix (8) constructed in Example 1. Its eigenvalues are equal when $\rho = 0$ and undergo a MPS when $\rho \in [0, 1)$ rises. So, by Theorem 1, the value Π of conceptual knowledge equals zero when $\rho = 0$ and is non-decreasing in ρ. Moreover, by Theorem 2, there is a threshold $\tau' \geq 0$ such that Π is increasing in τ if and only if $\tau < \tau'$. We characterize this threshold below. It equals zero when $\rho = 0$ and rises when ρ rises, consistent with Theorem 2.

Proposition 3. *Suppose θ has prior variance* (8) *with $\sigma^2 > 0$ and $\rho \in [0, 1)$. Then* Π

(i) equals zero when $\rho = 0$,

(ii) is increasing in ρ, and

(iii) is increasing in τ if and only if

$$\tau < \frac{\rho K}{1 + \rho(K-1)}.$$

Whereas Theorem 1 implies Π is non-decreasing in ρ, Proposition 3 says Π is increasing in ρ. This is because Theorem 1 holds for an arbitrary MPS, which may not affect the largest R^* eigenvalues and thus may not change the value (20) of an optimal sample. But this is impossible for the MPS induced by raising ρ, which raises the largest eigenvalue $\lambda_1 = (1 + \rho(K-1))\sigma^2$ of (8).

6 Deeper conceptual knowledge

Next, we study the value of "deepening" the agent's conceptual knowledge. We model this process as follows. Suppose the agent knows the trace

$$\text{tr}(\Sigma) = \sum_{k=1}^{K} \lambda_k$$

of the true prior variance matrix Σ and its first $J \in \{0, 1, \ldots, K\}$ eigenpairs (λ_k, v_k), but does not know the last $(K - J)$ eigenpairs. Intuitively, he knows the J concepts with the most explanatory power, but does not know the $(K - J)$ concepts with the least explanatory power. He assumes θ has prior variance

$$\Sigma^{(J)} = \sum_{k \leq J} \lambda_k v_k v_k^T + \lambda_K^{(J)} \left(I_K - \sum_{k \leq J} v_k v_k^T \right), \tag{25}$$

where

$$\lambda_K^{(J)} \equiv \frac{1}{K - J} \sum_{k > J} \lambda_k$$

is the mean of the smallest $(K-J)$ eigenvalues of Σ.[34] The matrix (25) has the same trace as Σ but (possibly) different eigenvalues; its k^{th} largest eigenvalue

$$\lambda_k^{(J)} \equiv \begin{cases} \lambda_k & \text{if } k \leq J \\ \lambda_K^{(J)} & \text{if } k > J \end{cases}$$

equals that of Σ if and only if $k \leq J$. The eigenvalues of $\Sigma^{(J)}$ have mean

$$\frac{1}{K}\sum_{k=1}^{K} \lambda_k^{(J)} = \overline{\lambda}$$

independently of J. Likewise $\lambda_K^{(0)} = \overline{\lambda}$ by definition. Thus $\Sigma^{(0)} = \overline{\lambda} I_K$ is the naïve prior variance matrix discussed in Sections 3.2 and 5. The parameter J interpolates between $\Sigma^{(0)}$ and $\Sigma^{(K)} = \Sigma$. It captures the "depth" of the agent's conceptual knowledge: the larger is J, the more concepts he knows and the richer is his mental model of θ.[35,36]

We say the agent has "J-deep conceptual knowledge" if his prior on θ has variance $\Sigma^{(J)}$. Suppose he has such knowledge and designs an optimal sample. Then, by analogy to (19) and (20), this sample has rank

$$R^{(J)} \equiv \max\left\{ k \in \{1, \ldots, K\} : \overline{\lambda}\left(\frac{k}{\lambda_k^{(J)}} - \sum_{j=1}^{k} \frac{1}{\lambda_j^{(J)}} \right) \leq \tau \right\}$$

and value

$$\pi^{(J)} \equiv \frac{\overline{\lambda}}{K}\left(\sum_{k=1}^{R^{(J)}} \frac{\lambda_k^{(J)}}{\overline{\lambda}} - \left(R^{(J)}\right)^2 \left(\sum_{k=1}^{R^{(J)}} \frac{\overline{\lambda}}{\lambda_k^{(J)}} + \tau \right)^{-1} \right),$$

where $\tau \equiv n\overline{\lambda}/\sigma_u^2$ is the precision index defined in Section 5. For example, letting $J = 0$ yields the rank $R^{(0)} = K$ and value

$$\pi^{(0)} = \frac{\overline{\lambda}\tau}{K + \tau}$$

of an optimal sample collected by a naïve agent. We refer to the difference

$$\Pi^{(J)} \equiv \pi^{(J)} - \pi^{(0)}$$

between $\pi^{(J)}$ and $\pi^{(0)}$ as the "value of J-deep conceptual knowledge." We characterize the relationship between $R^{(J)}$ and J in Lemma 4, and the relationship between $\Pi^{(J)}$ and J in Theorem 3.

[34] We define $\lambda_K^{(K)} \equiv \lambda_K$.

[35] Since $\lambda_1 \geq \cdots \geq \lambda_K$ (by assumption), there are non-increasing returns to knowing more concepts (i.e., increasing J), since each additional concept contributes a non-increasing share of states' prior variances. Intuitively, when the agent acquires conceptual knowledge, he prioritizes concepts with more explanatory power. For example, he takes classes or reads textbooks that provide "high-level summaries" before "digging into the details."

[36] For example, a farmer could deepen his conceptual knowledge by learning about different nutrients that crops need and fertilizers supply (e.g., nitrogen, phosphorus, and potassium). He could also learn about the nuances of nutrients he already knows about. Nutrients can take different chemical forms (e.g., nitrogen can take the form of ammonium and nitrate) that have different effects when applied to different crops. Learning about these differences gives the farmer new concepts that explain fertilizers' overall effects.

Lemma 4. *There is a threshold $J' \in \{0, \ldots, K\}$ such that*

$$R^{(J)} = \begin{cases} K & \text{if } J \leq J' \\ J & \text{if } J' < J < R^* \\ R^* & \text{if } J \geq R^* \end{cases}$$

for each $J \in \{0, \ldots, K\}$. This threshold is non-decreasing in τ.

Theorem 3. *The value $\Pi^{(J)}$ of J-deep conceptual knowledge*

(i) is non-negative,

(ii) equals zero when $J = 0$,

(iii) is non-decreasing in J, and

(iv) equals Π when $J \geq R^*$.

Theorem 3 says that deeper knowledge is (weakly) more valuable. Intuitively, knowing more concepts allows the agent to design samples that provide more payoff-relevant information.

The value of J-deep conceptual knowledge is bounded above by the value $\Pi^{(K)} = \Pi$ of "full" knowledge, and attains this bound when $J \geq R^*$. Thus, the agent gains no additional value from knowing more than the R^* concepts with the most explanatory power. This is because he ignores the other $(K - R^*)$ concepts when he designs samples (since $\delta_k^* = 0$ for each $k > R^*$), so knowing those concepts does not change his optimal sample.

For example, suppose the true prior variance matrix Σ has k^{th} largest eigenvalue

$$\lambda_k = \frac{K\alpha(1-\alpha)^{k-1}}{1-(1-\alpha)^K}$$

with $0 < \alpha < 1$. Then $\lambda_1, \ldots, \lambda_K$ are strictly positive, have mean $\overline{\lambda} = 1$, are constant in the limit as $\alpha \to 0$, and undergo a MPS as α rises.[37] This parameter determines the rate

$$\frac{\lambda_{k+1} - \lambda_k}{\lambda_k} = -\alpha$$

at which λ_k decays as k grows. Intuitively, the larger is α, the faster concepts' marginal explanatory power falls. Thus, if α is larger, then states are more reducible.

[37] For each $k \in \{1, \ldots, K\}$ we have $\lambda_k \to 1$ as $\alpha \to 0$ by L'Hôpital's rule. Moreover, the partial sum

$$\sum_{j=1}^{k} \lambda_k = \frac{K\Big(1-(1-\alpha)^k\Big)}{1-(1-\alpha)^K}$$

is non-decreasing in α and is constant in α when $k = K$. Thus, by Lemma 3, the eigenvalues $\lambda_1, \ldots, \lambda_K$ undergo a MPS when α rises.

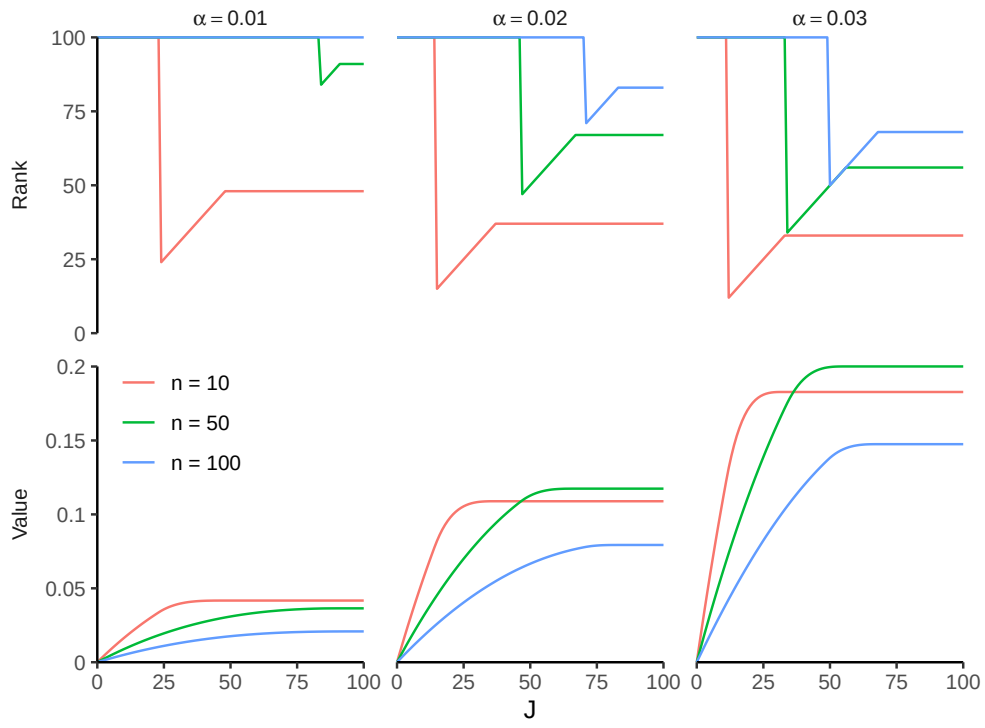

Figure 2: Rank $R^{(J)}$ and value $\Pi^{(J)}$ when $\lambda_{k+1} = (1-\alpha)\lambda_k$ and $(K, \overline{\lambda}, \sigma_u^2) = (100, 1, 1)$

Figure 2 shows how $R^{(J)}$ and $\Pi^{(J)}$ depend on J when $(K, \overline{\lambda}, \sigma_u^2) = (100, 1, 1)$ and λ_k decays at rate $\alpha \in \{0.01, 0.02, 0.03\}$. If J is sufficiently small, then the agent designs an optimal sample with rank $R^{(J)} = 100$; otherwise, he designs a sample with rank $R^{(J)} = \min\{J, R^*\}$. The threshold depth at which he switches from 100 to $\min\{J, R^*\}$ rises as the sample size n rises, consistent with Lemma 4.[38] The value $\Pi^{(J)}$ of J-deep conceptual knowledge is increasing in J when $J < R^*$ and constant in J when $J \geq R^*$. This value is increasing in α, consistent with Theorem 1: conceptual knowledge is more valuable when $\lambda_1, \ldots, \lambda_K$ are more spread out. Likewise $\Pi^{(J)}$ is non-monotone in n, consistent with Theorem 2: raising n allows the agent to learn more about the "in-sample" coefficients $\gamma_1, \ldots, \gamma_{R^{(J)}}$ (raising $\Pi^{(J)}$), but also prompts him to expand his sample and learn about more coefficients (lowering $\Pi^{(J)}$).

7 More concepts or more data?

Finally, we study the trade-off between conceptual and statistical knowledge. Our model in Section 6 provides a formal language for describing this trade-off: would the agent rather know more concepts (i.e., increase the depth J) or have more data (i.e., increase the sample size n)?[39]

[38] Here $\tau = n$ because $\overline{\lambda} = \sigma_u^2 = 1$.

[39] In contrast, Dominitz and Manski (2017) study the trade-off between having more data and "better" data.

Suppose the agent has J-deep conceptual knowledge and designs an optimal sample of size n. The value $\pi^{(J)}$ of this sample indexes the agent's welfare: it is larger when his minimized posterior expected loss is smaller. Lemma 5 says the agent is better off with deeper knowledge or more data. Intuitively, if he knows more concepts, then he can "ask better questions." If he has more data, then he can obtain better answers to his questions, making his posterior beliefs more precise and his expected loss lower.

Lemma 5. *The value $\pi^{(J)}$ is*

(i) non-decreasing in J and

(ii) increasing in n.

Now suppose the agent has a target value $\pi_0 \geq 0$. Let

$$n_{\pi_0}^{(J)} \equiv \min\{n \geq 0 : \pi^{(J)} \geq \pi_0\}$$

be the minimum sample size necessary to attain this value. This size is smaller when the agent knows more concepts and when states are more reducible:

Theorem 4. *Fix $\pi_0 \geq 0$. Then $n_{\pi_0}^{(J)}$*

(i) is non-increasing in J and

(ii) does not rise when $\lambda_1, \ldots, \lambda_K$ undergo a MPS.

Theorem 4 says that if the agent knows more concepts, then he can attain the same welfare with fewer observations, especially when states are highly reducible. This is because he can design better samples and extract more value from each observation, lowering the number he needs to attain the target π_0.

As in illustration of Theorem 4, suppose the true eigenvalues $\lambda_1, \ldots, \lambda_K$ have mean $\overline{\lambda} = 1$ and decay at rate α as in Section 6. Figure 3 shows how the minimum sample size $n_{\pi_0}^{(J)}$ depends on the welfare target π_0 and depth J when $(K, \sigma_u^2) = (100, 1)$ and $\alpha \in \{0.01, 0.02, 0.03\}$. Given π_0, the size $n_{\pi_0}^{(J)}$ is decreasing in J when $J < R^*$ and constant in J when $J \geq R^*$. Intuitively, if the agent knows too few concepts, then he cannot design samples that focus on all of the "right" concepts. Giving him more concepts empowers him to design better samples, extract more value from each observation, and require fewer observations to attain π_0. However, once he knows all the "right" concepts, giving him more does not change how he designs samples or the marginal value of each observation. Then the only way to for him to obtain *more valuable* data is to obtain *more* data, thus making $n_{\pi_0}^{(J)}$ constant in $J \geq R^*$.

The curves in Figure 3 are indifference curves: they trace out sets of depth-size pairs (J, n) that allow the agent to attain different welfare targets π_0. The slope of each curve equals the marginal rate of substitution (hereafter "MRS") between concepts and data. Intuitively, this MRS captures the number of observations the agent would give up to know another concept. It depends on the

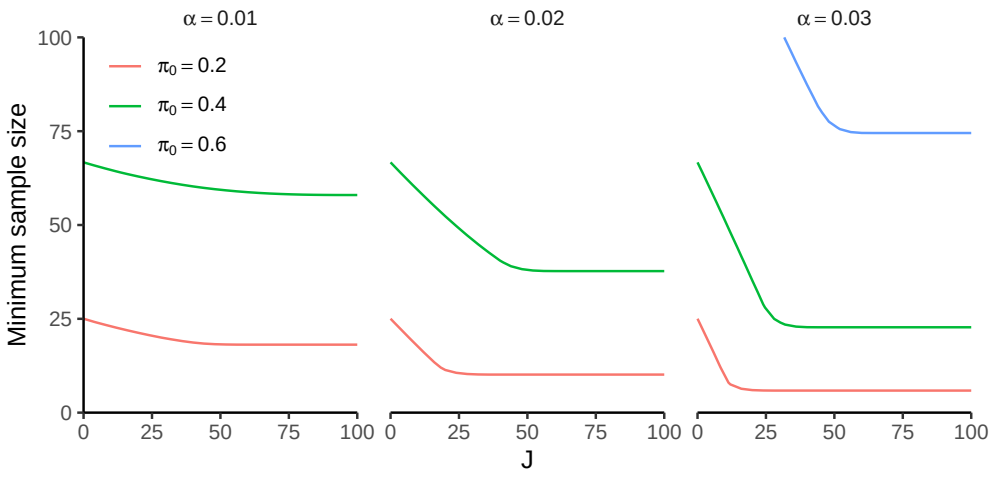

Figure 3: Minimum sample sizes $n_{\pi_0}^{(J)}$ when $\lambda_{k+1} = (1-\alpha)\lambda_k$ and $(K, \overline{\lambda}, \sigma_u^2) = (100, 1, 1)$

depth-size pair (J, n), the target π_0, and the parameter α indexing states' reducibility. Raising this parameter raises the rank R^* of the optimal sample he would design if he had full knowledge (see Figure 2). So raising α can have three effects on the MRS between concepts and data:

1. If $J < R^*$ before and after α rises, then the MRS rises in absolute value;
2. If $J < R^*$ before α rises but $J \geq R^*$ after, then the MRS falls in absolute value (to zero);
3. If $J \geq R^*$ before and after α rises, then the MRS remains unchanged (at zero).

So if states are more reducible, then the MRS between concepts and data may be higher or lower, depending on how many concepts the agent already knows. We defer analyzing this dependence to future research.

8 Modeling assumptions and related literature

We assume states and outcomes are jointly normally distributed under the agent's prior, and his actions are real-valued and induce quadratic losses. This assumption allows us to derive a closed-form expression for the value of conceptual knowledge. It is common in the literature on statistical decisions (Hastie et al., 2009). It is also implicit in empirical economics papers that estimate linear models via Ordinary Least Squares. Our agent has a linear model (5) of the unknown state vector. Moreover, if his prior is diffuse, then his optimal actions equal the estimates obtained via OLS.

We also assume the agent knows how the states covary *a priori*. This separates the conceptual knowledge embedded in his prior from the statistical knowledge he infers from data. It allows us to measure his conceptual and statistical knowledge on independent scales (i.e., depth and sample size), and to study their relative contributions to his welfare.

These assumptions impose a lot of structure on the agent's decision problem. However, they also give us a tractable and transparent framework for studying how conceptual and statistical knowledge interact. For example, they help us isolate and quantify the benefit of having conceptual knowledge: it empowers the agent to design more informative samples. Our framework also clarifies the mechanism through which samples become more informative and valuable: spreading the eigenvalues of the prior variance matrix focuses the agent on fewer dimensions. Our paper complements others on abstract decision problems (e.g., Blackwell, 1951, 1953; Brooks et al., 2024; Frankel and Kamenica, 2019; Whitmeyer, 2024), which offer different perspectives on what makes information more valuable.[40]

Our paper's premise—that conceptual knowledge is valuable—is intuitive and important, which is why it has been recognized by other economists: Lucas (1976) and Wolpin (2013) discuss how conceptual knowledge is essential for policy evaluation; Armstrong et al. (2025) and Fessler and Kasy (2019) discuss how it can be used to improve econometric estimators; Jackson (2019) discusses how it can guide experimental designs. Our contribution is to define and characterize the value of conceptual knowledge when making statistical decisions.

Our distinction between conceptual and statistical knowledge also appears in Ilut and Valchev (2025). They consider an agent who learns via "abstract reasoning" and "integrating experience." These learning modes correspond to our notions of conceptual and statistical knowledge. Ilut and Valchev study a dynamic setting, and focus on the "learning traps" that arise from reasoning too little or having the wrong data. In contrast, we study a static setting, and focus on the benefits of reasoning correctly and having the "right" data.

Ilut and Valchev model the cognitive processes that humans use to learn and make decisions. These processes differ from those used by machines: whereas humans can use concepts and causal reasoning, machines currently cannot (without human supervision). Instead they rely on pattern recognition and data-driven prediction (Felin and Holweg, 2024).

This difference between humans and machines motivates papers comparing their predictive performance (e.g., Kleinberg et al., 2018; Kühl et al., 2022; Mullainathan and Obermeyer, 2022). We shed light on *when* humans are likely to outperform "naïve" (in the sense defined in Section 3.2) machines: when parameters of interest are highly reducible (see Theorem 1), when data are noisy or scarce (see Theorem 2), and when the sampling frame is limited (see Appendix Section A2.3). We also shed light on *why* humans can outperform machines: humans have conceptual knowledge that helps us "ask the right questions."

Whereas we compare humans and machines implicitly, Iakovlev and Liang (2025) compare them explicitly. They study the "value of context": how much predictive power one gains from choosing the "right" covariates, relative to an algorithm that cannot make this choice. They show that the value of context vanishes as one's access to data becomes large. Similarly, we show that the value of conceptual knowledge vanishes as one's access to data becomes large (see Theorem 2).

[40]Other authors study the value of information in specific decisions problems different than that faced by our agent (e.g., Athey and Levin, 2018; Cabrales et al., 2013; Lehmann, 1988; Moscarini and Smith, 2002; Persico, 2000).

Together, these results suggest that conceptual knowledge and data are substitutes in the "big data worlds" ruled by machines, which can detect patterns in data without knowing why those patterns arise. But Theorem 2 also suggests that concepts and data are complements in the "small data worlds" ruled by humans: we can use our conceptual knowledge to learn more from small samples. Moreover, Theorem 2 draws the border (the threshold τ') between these big and small data worlds.

Our paper also connects to the literature on model-based learning. Andrews et al. (2025) find that "black box" algorithms outperform models when predicting within domains, but are worse at generalizing across domains.[41] This is consistent with our idea that models embed conceptual knowledge (see Appendix Section A1.2) and that such knowledge boosts out-of-sample predictive performance (see Appendix Section A2.3). Fudenberg et al. (2022) propose a measure of model "completeness": the share of reducible prediction error that imposing a model reduces. Our paper suggests a different notion of completeness: a model is "more complete" when it embeds deeper conceptual knowledge.[42]

In our analysis of deeper conceptual knowledge (in Sections 6 and 7), we assume there is a true model of the agent's environment that he can know at different depths. This is in contrast to the literature on model uncertainty and mis-specification, which considers settings where the true model is unknown (Cerreia-Vioglio et al., 2025; Chatfield, 1995; Esponda and Pouzo, 2016; Hansen and Sargent, 2001; Marinacci, 2015). In these settings, many authors advocate a "robust" approach that imposes minimal structural restrictions and provides payoff guarantees across a range of possible models (Gilboa and Schmeidler, 1989; Klibanoff et al., 2005). This is consistent with our "naïve" agent, who lacks conceptual knowledge, assuming a minimally restrictive prior. We add to the model uncertainty and mis-specification literature by quantifying the value of knowing and imposing the correct structural restrictions.

We also assume away any competitive or political forces that may lead the agent to choose one model over another (Dasaratha et al., 2025; Izzo et al., 2023). These forces arise in the literature on "models as narratives" (Aina, 2025; Schwartzstein and Sunderam, 2021; Eliaz et al., 2025). Papers in that literature focus on settings where many models are plausibly true (e.g., financial markets and political campaigns). In contrast, many real-world decisions are made in settings where there is a single, objectively true model that can be learned through education and introspection.[43] Such settings are our focus in this paper. For this reason, we do not consider the issues that would arise if the agent's conceptual knowledge was mis-specified or supplied by a strategic communicator

[41] See also Fudenberg and Liang (2019), Peterson et al. (2021), and Peysakhovich and Naecker (2017) for comparisons of model-based and black box predictions.

[42] Mailath and Samuelson (2020) argue that "in practice, people work with models that are deliberately incomplete, including the most salient variables and excluding others." Indeed, Theorem 3 implies that the agent does not benefit from using models embedding depths greater than the rank R^* of an optimal sample.

[43] For example, there exists an objectively true set of mechanisms through which fertilizers help crops grow. A farmer could discover these mechanisms experimentally or be taught them (see Sankar et al., 2025), or may rely on heuristics that capture some mechanisms but not all (e.g., the farmer may notice that two fertilizers intensify greening, but not know they do so because they supply nitrogen).

with competing incentives. However, we believe these issues are interesting and worthy of future research.

9 Conclusion

This paper introduces a simple idea: whereas information is valuable because it leads to better decisions, conceptual knowledge is valuable because it leads to better information. We formalize this idea and study its consequences. Conceptual knowledge is more valuable when payoff-relevant unknowns are more "reducible": when they can be explained with fewer common concepts. Its value is non-monotone in the quantity and quality of available data, and vanishes with infinite data. Deeper conceptual knowledge is (weakly) more valuable and compensates for having less data. So we can improve interventions that give people data by giving them conceptual tools to interpret data. We can also improve human-AI interactions by recognizing humans' comparative advantage: knowing how to "ask the right questions."

Stemming from this paper are several avenues for future research. One is to analyze the trade-off between concepts and data in a consumer choice setting. This would require specifying the "price" of acquiring concepts vis-à-vis observations of data. Given a price schedule, one could ask many questions about concepts and data: Are they complements or substitutes? Are they normal or inferior? How do these statuses depend on the states' reducibility?

Another avenue is to make our framework dynamic. For example, the agent could take actions that generate outcomes observed by future agents. This would allow us to study how conceptual knowledge is "discovered" and passed down to new generations. This discovery process has been studied by other authors (Carnehl and Schneider, 2025; Gans, 2025); blending their models and insights with ours may bear fruit.

A third avenue is to study the role of conceptual knowledge in social learning and technology adoption. This paper and its empirical sibling (Sankar et al., 2025) came from studying agricultural settings, where heterogeneity in outcomes and the inability to generalize slows learning (Conley and Udry, 2010; Munshi, 2004; Tjernström, 2017) and technology adoption (Alidaee, 2023; BenYishay and Mobarak, 2019; Laajaj and Macours, 2024). These frictions could be removed by giving people conceptual tools that enable them to generalize. This paper offers a framework for studying such tools and their value.

References

Aina, C. (2025). Tailored Stories.

Alidaee, H. (2023). How Uncertainty About Heterogeneity Impacts Technology Adoption.

Andrews, I., Fudenberg, D., Lei, L., Liang, A., and Wu, C. (2025). The Transfer Performance of Economic Models.

Armstrong, T. B., Kline, P., and Sun, L. (2025). Adapting to Misspecification. *Econometrica*, forthcoming.

Arnold, B. C. (1987). *Majorization and the Lorenz Order: A Brief Introduction*, volume 43 of *Lecture Notes in Statistics*. Springer New York, New York, NY.

Arrow, K. J. (1963). Uncertainty and the Welfare Economics of Medical Care. *American Economic Review*, 53(5):941–973.

Athey, S. and Levin, J. (2018). The value of information in monotone decision problems. *Research in Economics*, 72(1):101–116.

Bardhi, A. (2024). Attributes: Selective Learning and Influence. *Econometrica*, 92(2):311–353.

Bartlett, M. S. (1951). An Inverse Matrix Adjustment Arising in Discriminant Analysis. *Annals of Mathematical Statistics*, 22(1):107–111.

BenYishay, A. and Mobarak, A. M. (2019). Social Learning and Incentives for Experimentation and Communication. *Review of Economic Studies*, 86(3):976–1009.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Blackwell, D. (1951). Comparison of Experiments. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press.

Blackwell, D. (1953). Equivalent Comparisons of Experiments. *Annals of Mathematical Statistics*, 24(2):265–272.

Brooks, B., Frankel, A., and Kamenica, E. (2024). Comparisons of Signals. *American Economic Review*, 114(9):2981–3006.

Cabrales, A., Gossner, O., and Serrano, R. (2013). Entropy and the Value of Information for Investors. *American Economic Review*, 103(1):360–377.

Callander, S. (2011). Searching and Learning by Trial and Error. *American Economic Review*, 101(6):2277–2308.

Carnehl, C. and Schneider, J. (2025). A Quest for Knowledge. *Econometrica*, 93(2):623–659.

Cerreia-Vioglio, S., Hansen, L. P., Maccheroni, F., and Marinacci, M. (2025). Making Decisions Under Model Misspecification. *Review of Economic Studies*, forthcoming.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419.

Conley, T. G. and Udry, C. R. (2010). Learning about a New Technology: Pineapple in Ghana. *American Economic Review*, 100(1):35–69.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, second edition.

Dasaratha, K., Ortner, J., and Zhu, C. (2025). Markets for Models.

Davies, B. (2024). Learning about a changing state.

DeGroot, M. H. (2004). *Optimal Statistical Decisions*. Wiley, first edition.

Dominitz, J. and Manski, C. F. (2017). More Data or Better Data? A Statistical Decision Problem. *Review of Economic Studies*, 84(4):1583–1605.

Eliaz, K., Galperti, S., and Spiegler, R. (2025). False Narratives and Political Mobilization. *Journal of the European Economic Association*, 23(3):983–1027.

Esponda, I. and Pouzo, D. (2016). Berk-Nash Equilibrium: A Framework for Modeling Agents With Misspecified Models. *Econometrica*, 84(3):1093–1130.

Felin, T. and Holweg, M. (2024). Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. *Strategy Science*, 9(4):346–371.

Fessler, P. and Kasy, M. (2019). How to Use Economic Theory to Improve Estimators: Shrinking Toward Theoretical Restrictions. *Review of Economics and Statistics*, 101(4):681–698.

Fortiana, J. and Cuadras, C. (1997). A family of matrices, the discretized Brownian bridge, and distance-based regression. *Linear Algebra and its Applications*, 264:173–188.

Frankel, A. and Kamenica, E. (2019). Quantifying Information and Uncertainty. *American Economic Review*, 109(10):3650–3680.

Fudenberg, D., Kleinberg, J., Liang, A., and Mullainathan, S. (2022). Measuring the Completeness of Economic Models. *Journal of Political Economy*, 130(4):956–990.

Fudenberg, D. and Liang, A. (2019). Predicting and Understanding Initial Play. *American Economic Review*, 109(12):4112–4141.

Gans, J. S. (2025). A Quest for AI Knowledge. NBER Working Paper w33566, National Bureau of Economic Research, Cambridge, MA.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153.

Gollier, C. and Schlesinger, H. (1996). Arrow's theorem on the optimality of deductibles: A stochastic dominance approach. *Economic Theory*, 7(2):359–363.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. Adaptive computation and machine learning. MIT press, Cambridge, Mass.

Halevy, A., Norvig, P., and Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2):8–12.

Hansen, L. P. and Sargent, T. J. (2001). Robust Control and Model Uncertainty. *American Economic Review*, 91(2):60–66.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, second edition.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, New York, NY, second edition.

Howard, R. (1966). Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26.

Iakovlev, A. and Liang, A. (2025). The Value of Context: Human versus Black Box Evaluators.

Ilut, C. and Valchev, R. (2025). Learning Optimal Behavior Through Reasoning and Experiences.

Izzo, F., Martin, G. J., and Callander, S. (2023). Ideological Competition. *American Journal of Political Science*, 67(3):687–700.

Jackson, M. O. (2019). The Role of Theory in an Age of Design and Big Data. In Laslier, J.-F., Moulin, H., Sanver, M. R., and Zwicker, W. S., editors, *The Future of Economic Design*, pages 523–530. Springer International Publishing, Cham.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Number 6 in Cognitive science series. Harvard University Press, Cambridge, MA.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics*, 133(1):237–293.

Klibanoff, P., Marinacci, M., and Mukerji, S. (2005). A Smooth Model of Decision Making under Ambiguity. *Econometrica*, 73(6):1849–1892.

Kühl, N., Goutier, M., Baier, L., Wolff, C., and Martin, D. (2022). Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, 76:78–92.

Laajaj, R. and Macours, K. (2024). The Complexity of Multidimensional Learning in Agriculture. CEPR Discussion Paper DP19009, CEPR Press, Paris and London.

Lehmann, E. L. (1988). Comparing Location Experiments. *Annals of Statistics*, 16(2).

Levin, D., Peres, Y., and Wilmer, E. (2008). *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island.

Liang, A., Mu, X., and Syrgkanis, V. (2022). Dynamically Aggregating Diverse Information. *Econometrica*, 90(1):47–80.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

Mailath, G. J. and Samuelson, L. (2020). Learning under Diverse World Views: Model-Based Inference. *American Economic Review*, 110(5):1464–1501.

Marinacci, M. (2015). Model Uncertainty. *Journal of the European Economic Association*, 13(6):1022–1100.

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, 1505(1):79–101.

Moscarini, G. and Smith, L. (2002). The Law of Large Demand for Information. *Econometrica*, 70(6):2351–2366.

Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care. *Quarterly Journal of Economics*, 137(2):679–727.

Munshi, K. (2004). Social learning in a heterogeneous population: Technology diffusion in the Indian Green Revolution. *Journal of Development Economics*, 73(1):185–213.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press.

Persico, N. (2000). Information Acquisition in Auctions. *Econometrica*, 68(1):135–148.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., and Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214.

Peysakhovich, A. and Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge, MA.

Rothschild, M. and Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, 2(3):225–243.

Sankar, A., Dulin, R., Davies, B., Nourani, V., Rudder, J., Salomon, A., and Taulya, G. (2025). How mechanistic explanations reshape learning and behavior: Evidence from a fertilizer choice experiment in Eastern Uganda.

Schwartzstein, J. (2014). Selective Attention and Learning. *Journal of the European Economic Association*, 12(6):1423–1452.

Schwartzstein, J. and Sunderam, A. (2021). Using Models to Persuade. *American Economic Review*, 111(1):276–323.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285.

Tjernström, E. (2017). Learning from Others in Heterogeneous Environments.

Varian, H. R. (1987). The Arbitrage Principle in Financial Economics. *Journal of Economic Perspectives*, 1(2):55–72.

Whitmeyer, M. (2024). Making Information More Valuable. *Journal of Political Economy*, forthcoming.

Wolpin, K. I. (2013). *The Limits of Inference without Theory*. Tjalling C. Koopmans Memorial Lectures. MIT Press.

Yu, Y., Wang, T., and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323.

A Additional material

A1 Connection to statistical learning

This section connects our paper to the literatures on machine and statistical learning, which study how to derive predictive functions from data.[44] First, we show that our framework (described in Section 3) can be used to study Bayesian learning about real-valued functions. Second, we show how the agent's prior derives from his approximating model of an unknown function.

A1.1 Function-state equivalence

Suppose there is a finite set $\mathcal{X}$ of "inputs" and a square-summable function $f : \mathcal{X} \to \mathbb{R}$ belonging to the set

$$\mathcal{F} \equiv \left\{ g \in \mathbb{R}^{\mathcal{X}} : \sum_{x \in \mathcal{X}} (g(x))^2 < \infty \right\}$$

of such functions. Endow $\mathcal{F}$ with the inner product defined by

$$\langle g, g' \rangle = \sum_{x \in \mathcal{X}} g(x) g'(x)$$

for all pairs $(g, g') \in \mathcal{F} \times \mathcal{F}$. Let $|\mathcal{X}| = K$, let $\ell : \mathcal{X} \to \{1, \ldots, K\}$ be a bijection, and define

$$\phi_k(x) \equiv \begin{cases} 1 & \text{if } \ell(x) = k \\ 0 & \text{otherwise} \end{cases}$$

for each $x \in \mathcal{X}$ and $k \in \{1, \ldots, K\}$. Then the indicator functions $\phi_1, \ldots, \phi_K$ form an orthonormal basis $\mathcal{B} \equiv \{\phi_k\}_{k=1}^K$ for the inner product space $(\mathcal{F}, \langle ., . \rangle)$. Now let $\theta_1, \ldots, \theta_K$ be the coordinates of f over $\{\phi_k\}_{k=1}^K$:

$$f = \sum_{k=1}^{K} \theta_k \phi_k.$$

The agent knows $\phi_1, \ldots, \phi_K$ but not $\theta \equiv (\theta_1, \ldots, \theta_K)$, so learning about f is equivalent to learning about θ.[45] Moreover, suppose the agent draws an input $x \in \mathcal{X}$ uniformly at random, and predicts the "output" $y \in \mathbb{R}$ with conditional distribution

$$y \mid x, f \sim \mathcal{N}\big(f(x), \sigma_u^2\big)$$

given x and f. His "prediction rule" $\hat{f} \in \mathcal{F}$ maps each realization of x to a prediction $\hat{f}(x)$ of y. This prediction induces a posterior mean squared error (MSE)

$$\mathbb{E}\Big[(y - \hat{f}(x))^2 \mid x, \mathcal{S}\Big] = \mathbb{E}\Big[(f(x) - \hat{f}(x))^2 \mid x, \mathcal{S}\Big] + \sigma_u^2$$

[44] See Bishop (2006) or Hastie et al. (2009) for textbook treatments.

[45] Assuming θ is normally distributed is equivalent to assuming $\{f(x)\}_{x \in \mathcal{X}}$ follows a Gaussian process. Such processes arise in the economic literature on learning and information acquisition—see, e.g., Bardhi (2024), Davies (2024), Ilut and Valchev (2025), or Laajaj and Macours (2024). See also Bishop (2006, Section 6.4) or Rasmussen and Williams (2006) for more information about Gaussian processes and their applications.

given x and the sample $\mathcal{S}$, where $\mathbb{E}$ takes expectations with respect to the joint prior distribution of input-output pairs. The agent chooses $\hat{f}$ to minimize the mean posterior MSE across realizations of x:

$$\hat{f} \in \arg\min_{g \in \mathcal{F}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{E}\Big[(y - g(x))^2 \mid x, \mathcal{S}\Big]. \tag{A1}$$

The optimal actions $a_1, \ldots, a_K$ defined by (3) are precisely the coordinates of $\hat{f}$ over the basis $\mathcal{B}$. By Lemma 3, these coordinates equal the posterior mean coordinates of f. The minimized mean posterior MSE

$$\min_{g \in \mathcal{F}} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{E}\Big[(y - \hat{f}(x))^2 \mid x, \mathcal{S}\Big] = \mathbb{E}[L(\theta, a) \mid \mathcal{S}] + \sigma_u^2$$

equals the expected loss (10) plus a constant σ_u^2 that arises due to the irreducible randomness in the outcome y. Thus, the prediction problem (A1) is equivalent to the choice problem (3).

A1.2 Approximating models

Suppose the agent knows about a collection $\psi_1, \ldots, \psi_J \in \mathcal{F}$ of "features" that (partially) mediate the relationship between inputs and outputs.[46] These features are linearly independent (but not necessary orthonormal) elements of the function space $\mathcal{F}$. They map inputs to known, measurable quantities. For example, if the inputs are fertilizers, then the features could map fertilizers to quantities of different nutrients.

The agent uses $\psi_1, \ldots, \psi_J$ to build a "model" $m \in \mathcal{F}$ approximating the unknown function f. This model is a linear combination of features: there is a(n unknown) vector $\beta \equiv (\beta_1, \ldots, \beta_J) \in \mathbb{R}^J$ such that

$$m(x) = \sum_{k=1}^{J} \beta_k \psi_k(x)$$

for each $x \in \mathcal{X}$. Then the derivative

$$\frac{\partial m(x)}{\partial \psi_k(x)} = \beta_k$$

of $m(x)$ with respect to $\psi_k(x)$ does not depend on the input x. In this way, the model m captures the generalizable structure of f that is common to all inputs. In contrast, the model's approximation error

$$\epsilon \equiv f - m$$

captures the idiosyncrasies specific to each input.

The agent uses his knowledge of $\psi_1, \ldots, \psi_J$ to construct his prior $\mathbb{P}$ on θ. First, he identifies the subspace

$$\mathcal{F}^m \equiv \text{span}\{\psi_1, \ldots, \psi_J\}$$

[46] This J is the same as the depth parameter defined in Section 6.

of $\mathcal{F}$ spanned by the features. It corresponds to a subspace

$$\Theta^m \equiv \left\{ \vartheta \in \Theta : \sum_{k=1}^{K} \vartheta_k \phi_k \in \mathcal{F}^m \right\}$$

of the Euclidean space $\Theta \equiv \mathbb{R}^K$ containing the unknown coordinate vector θ. Concretely, if $\theta^j \equiv (\theta^j_1, \ldots, \theta^j_K)$ contains the (known) coordinates of the j^{th} feature ψ_j over the orthonormal basis $\mathcal{B}$, then

$$\Theta^m = \text{span}\{\theta^1, \ldots, \theta^J\}$$

is the subspace of Θ spanned by the vectors $\theta^1, \ldots, \theta^J$.

Next, the agent constructs an orthonormal basis $\{v_k\}_{k=1}^J$ for Θ^m (e.g., by applying the Gram-Schmidt process to $\theta^1, \ldots, \theta^J$). If $J < K$, then he also constructs an orthonormal basis $\{v_k\}_{k=J+1}^K$ for the orthogonal complement

$$\Theta^\epsilon \equiv \left\{ \vartheta \in \Theta : \vartheta^T \vartheta' = 0 \text{ for all } \vartheta' \in \Theta^m \right\}$$

of Θ^m. Then $\{v_k\}_{k=1}^K$ is an orthonormal basis for Θ. Letting $\gamma \equiv (\gamma_1, \ldots, \gamma_K)$ contain the coordinates of θ over $\{v_k\}_{k=1}^K$ yields (5); the eigendecomposition (6) of Σ follows. Thus, the agent's prior on θ derives from his prior on γ, which derives from his knowledge of $\psi_1, \ldots, \psi_J$ (which define the model m).

A2 Value of $\mathcal{S}$

Consider the sample $\mathcal{S}$. Proposition A1 says that the value $\pi(\mathcal{S})$ of $\mathcal{S}$ is non-negative, grows as $\mathcal{S}$ grows, and shrinks as σ_u^2 grows.

Proposition A1. *The value $\pi(\mathcal{S})$ of the sample $\mathcal{S}$*

(i) is non-negative,

(ii) does not fall when $\mathcal{S}$ gains observations, and

(iii) falls when σ_u^2 rises.

Sections A2.1–A2.3 discuss the values of samples with specific structures.

A2.1 Singleton samples

Suppose $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ contains a single observation. Then the Gram matrix $G = w^{(1)}(w^{(1)})^T$ has eigenvalues $\delta_1 = 1$ and $\delta_2 = \cdots = \delta_K = 0$. Substituting them into (17) gives us bounds on the value of $\mathcal{S}$:

Proposition A2. *Suppose $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ contains a single observation. Then its value*

$$\pi(\mathcal{S}) = \frac{(w^{(1)})^T \Sigma^2 w^{(1)}}{K\big((w^{(1)})^T \Sigma w^{(1)} + \sigma_u^2\big)} \tag{A2}$$

satisfies

$$\frac{\lambda_K^2}{K(\lambda_K + \sigma_u^2)} \overset{\star}{\leq} \pi(\mathcal{S}) \overset{\star\star}{\leq} \frac{\lambda_1^2}{K(\lambda_1 + \sigma_u^2)}, \tag{A3}$$

where $\star$ holds with equality if $\Sigma w^{(1)} = \lambda_K w^{(1)}$ and $\star\star$ holds with equality if $\Sigma w^{(1)} = \lambda_1 w^{(1)}$.

The value of $\{(w^{(1)}, y^{(1)})\}$ is largest when $w^{(1)}$ is an eigenvector of Σ with corresponding eigenvalue $\lambda_1 = \max\{\lambda_1, \ldots, \lambda_K\}$. It is smallest when $w^{(1)}$ is an eigenvector of Σ with corresponding eigenvalue $\lambda_K = \min\{\lambda_1, \ldots, \lambda_K\}$. Intuitively, the more "weight" $w^{(1)}$ puts on directions in which the prior variance of θ is large, the more valuable it is to observe $(w^{(1)}, y^{(1)})$ because the larger is the variance reduction it delivers. This is especially true when there are few dimensions (i.e., K is small) and when the signal $y^{(1)}$ is precise (i.e., σ_u^2 is small).

For example, suppose Σ is the matrix (8) constructed in Example 1. Let $K = 2$ and suppose $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ contains a single observation with

$$w^{(1)} = (\sin(\pi t), \cos(\pi t))$$

and $-1/2 \leq t \leq 1/2$. Increasing t from $-1/2$ to $1/2$ rotates $w^{(1)}$ clockwise from $(-1, 0)$ to $(1, 0)$. The value[47]

$$\pi(\mathcal{S}) = \frac{\big(1 + 2\rho \sin(2\pi t) + \rho^2\big)\sigma^4}{2((1 + \rho \sin(2\pi t))\sigma^2 + \sigma_u^2)} \tag{A4}$$

of $\mathcal{S}$ attains its minimum when $t = -1/4$, in which case $w^{(1)} = (-1/\sqrt{2}, 1/\sqrt{2})$ equals the unit eigenvector v_2 of Σ with the smallest corresponding eigenvalue. In contrast, the value of $\mathcal{S}$ attains its maximum when $t = 1/4$, in which case $w^{(1)} = (1/\sqrt{2}, 1/\sqrt{2})$ equals the unit eigenvector v_1 of Σ with the largest corresponding eigenvalue. Figure A1 shows that $\pi(\mathcal{S})$ rises monotonically as t rises from $-1/4$ to $1/4$, which lowers the angle between $w^{(1)}$ and v_1 from 90° to 0°.

A2.2 Representative samples

Suppose the covariates $w^{(1)}, \ldots, w^{(n)}$ in $\mathcal{S}$ are binary vectors: for each $i \in \{1, \ldots, n\}$, there is an index $k_i \in \{1, \ldots, K\}$ such that $w^{(i)}$ has k^{th} component

$$w_k^{(i)} = \begin{cases} 1 & \text{if } k = k_i \\ 0 & \text{otherwise.} \end{cases}$$

Then each outcome

$$y^{(i)} = \theta_{k_i} + u^{(i)}$$

[47] We obtain (A4) by substituting $K = 2$, the prior variance matrix (8), the covariate $w^{(1)} = (\sin(\pi t), \cos(\pi t))$, and the sample size $n = 1$ into (A2).

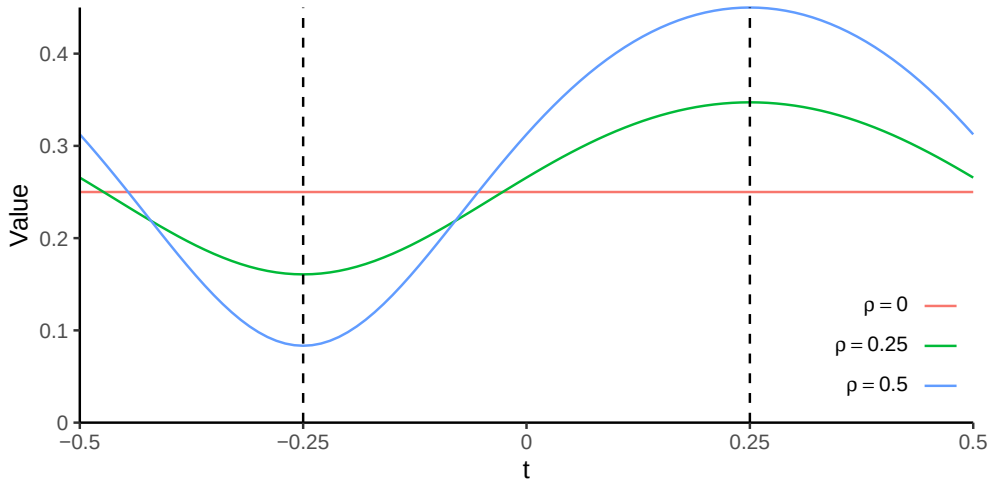

Figure A1: Value (A4) of observing $\mathcal{S} = \{((\sin(\pi t), \cos(\pi t)), y^{(1)})\}$ when θ has prior variance (8) and $(K, \sigma^2, \sigma_u^2) = (2, 1, 1)$

is a "pure signal" of the state θ_{k_i}. Moreover, the Gram matrix G is diagonal: its kk^{th} entry

$$G_{kk} = |\{i \in \{1, \ldots, K\} : k_i = k\}|$$

counts the outcomes in $\mathcal{S}$ that are pure signals of θ_k. If $G_{11}, \ldots, G_{KK}$ are equal (to n/K), then the eigenvalues of G are also equal (to n/K). These eigenvalues characterize a sample that contains equal information about each state.

Accordingly, we say $\mathcal{S}$ is "representative" if it induces a Gram matrix with equal eigenvalues. Then the lower and upper bounds in (17) are equal, and so $\mathcal{S}$ has value

$$\begin{aligned}\pi(\mathcal{S}) &= \frac{1}{K}\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{n}{K\sigma_u^2}\right)^{-1}\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\frac{n\lambda_k^2}{n\lambda_k + K\sigma_u^2}.\end{aligned}$$

This value is larger when the eigenvalues $\lambda_1, \ldots, \lambda_K$ of Σ are more spread out:

Proposition A3. *If $\mathcal{S}$ is representative, then its value $\pi(\mathcal{S})$ does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS.*

If $\mathcal{S}$ is representative, then it contains equal information about each component of θ. But there are diminishing returns to having more information about a given component. So if the prior variances of $\gamma_1, \ldots, \gamma_K$ change in a mean-preserving way, then the increased reduction of the higher variances more than offsets the decreased reduction of the lower variances, thereby raising $\pi(\mathcal{S})$.

For example, suppose Σ is the matrix (8) constructed in Example 1. Then the eigenvalues $\lambda_1 = (1 + \rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$ of Σ undergo a MPS when ρ rises. So if $\mathcal{S}$ is

representative, then its value

$$\pi(\mathcal{S}) = \sigma^2 - \frac{1}{K}\left(\frac{1}{\lambda_1} + \frac{n}{K\sigma_u^2}\right)^{-1} - \left(1 - \frac{1}{K}\right)\left(\frac{1}{\lambda_K} + \frac{n}{K\sigma_u^2}\right)^{-1}$$

must be non-decreasing in ρ. Indeed, the derivative

$$\frac{\partial \pi(\mathcal{S})}{\partial \rho} = \left(1 - \frac{1}{K}\right)\sigma^2\left(\frac{1}{\lambda_K^2} - \frac{1}{\lambda_1^2}\right)\left(1 + \frac{n}{K\sigma_u^2}\right)^{-2}$$

of $\pi(\mathcal{S})$ with respect to ρ is non-negative because $\lambda_K \geq \lambda_1$.

A2.3 Non-spanning samples

If the sample $\mathcal{S}$ is *not* representative, then its value can fall when the eigenvalues of Σ undergo a MPS. This happens, for example, when $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ is a singleton and $w^{(1)}$ is an eigenvector of Σ corresponding to an eigenvalue that falls under the MPS.[48] Such a sample is "non-spanning": the rank

$$R \equiv \max\{k \in \{1, \ldots, K\} : \delta_k > 0\}$$

of the Gram matrix G is strictly less than K, so there are components of θ about which $\mathcal{S}$ contains no information because they are outside the column space

$$\begin{aligned} \text{col}(G) &\equiv \text{span}\{\omega_1, \ldots, \omega_K\} \\ &= \text{span}\{w^{(1)}, \ldots, w^{(n)}\} \end{aligned}$$

of G. The agent cannot learn about these components from $\mathcal{S}$ directly. But he can learn about them indirectly if he knows how they covary with the components that belong to $\text{col}(G)$.

For example, suppose the observations in $\mathcal{S}$ are pure signals of $\theta_1, \ldots, \theta_R$. Then

$$\text{col}(G) = \left\{v \in \mathbb{R}^K : v_k = 0 \text{ for each } k > R\right\}$$

is the subspace of $\mathbb{R}^K$ spanned by the first R standard basis vectors. So the first R components of θ are "on-support" but the last $(K - R)$ components are "off-support." Let $\theta_{\mathcal{S}} \equiv (\theta_1, \ldots, \theta_R)$ contain the first R components of θ and define

$$\xi_k \equiv \mathbb{V}(\theta_{\mathcal{S}})^{-1}\begin{bmatrix} \Sigma_{11} \\ \vdots \\ \Sigma_{Rk} \end{bmatrix}$$

for each $k \in \{1, \ldots, K\}$.[49] Then the posterior variance matrix has trace

$$\text{tr}(\mathbb{V}(\theta \mid \mathcal{S})) = \underbrace{\sum_{k=1}^{R} \mathbb{V}(\theta_k \mid \mathcal{S}) + \sum_{k>R} \xi_k^T \mathbb{V}(\theta_{\mathcal{S}} \mid \mathcal{S})\xi_k}_{\text{Sampling error}} + \underbrace{\sum_{k>R} \mathbb{V}(\theta_k \mid \theta_{\mathcal{S}})}_{\text{Extrapolation error}}, \tag{A5}$$

[48] For example, if θ has prior variance (8), then the value of $\mathcal{S} = \{(w^{(1)}, y^{(1)})\}$ is decreasing in ρ when $w^{(1)} = v_2$.

[49] The matrix $\mathbb{V}(\theta_S)$ is invertible because it is a leading principal submatrix of an invertible matrix.

where

$$\mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) = \mathbb{V}(\theta_k) - \xi_k^T \, \mathbb{V}(\theta_{\mathcal{S}}) \xi_k \tag{A6}$$

is the prior variance of θ_k left unexplained $\theta_{\mathcal{S}}$.[50] The first two terms on the RHS of (A5) are "sampling errors" that depend on how much information $\mathcal{S}$ contains about the on-support components $\theta_1, \ldots, \theta_R$. The third term is an "extrapolation error" that depends on how much information these components contain about the off-support components $\theta_{R+1}, \ldots, \theta_K$. Whereas the sampling error can be reduced by collecting more (or less noisy) data on $\theta_1, \ldots, \theta_R$, the extrapolation error cannot. It can only be reduced by making the on-support components of θ contain more informatio about the off-support components.

For example, suppose Σ is the matrix (8) constructed in Example 1. Then

$$\mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) = \frac{(1-\rho)(1+\rho R)\sigma^2}{1+\rho(R-1)} \tag{A7}$$

for each $k > R$ and so the extrapolation error

$$\sum_{k>R} \mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) = \frac{(1-\rho)(1+\rho R)(K-R)\sigma^2}{1+\rho(R-1)}$$

falls as the correlation ρ rises. It equals $(K-R)\sigma^2$ when $\rho = 0$, in which case $\theta_1, \ldots, \theta_R$ provide no information about $\theta_{R+1}, \ldots, \theta_K$ and so $\mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) = \mathbb{V}(\theta_k) = \sigma^2$ for each $k > R$. It equals zero in the limit as $\rho \to 1$, in which case $\theta_1, \ldots, \theta_K$ are fully determined by the coefficient γ_1 on their common component and so $\mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) = \mathbb{V}(\theta_k \mid \gamma_1) = 0$ for each $k \in \{1, \ldots, K\}$.

If, in addition, the observations in $\mathcal{S}$ have no noise, then $\mathbb{V}(\theta_k \mid \mathcal{S}) = 0$ for each $k \leq R$ and so $\mathcal{S}$ has value

$$\begin{aligned}
\lim_{\sigma_u^2 \to 0} \pi(\mathcal{S}) &= \frac{1}{K}\left(\sum_{k=1}^{R}(\mathbb{V}(\theta_k) - 0) + \sum_{k>R}(\mathbb{V}(\theta_k) - \mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}))\right) \\
&= \left(1 - \frac{(1-\rho)(1+\rho R)(K-R)}{(1+\rho(R-1))K}\right)\sigma^2.
\end{aligned}$$

This value rises as ρ rises, equals $R\sigma^2/K$ when $\rho = 0$, and equals σ^2 in the limit as $\rho \to 1$. Taking the limit as $K \to \infty$ gives

$$\lim_{K \to \infty} \lim_{\sigma_u^2 \to 0} \pi(\mathcal{S}) = \frac{\rho^2 R \sigma^2}{(1+\rho(R-1))},$$

which is bounded away from zero if and only if $\rho > 0$. So if there are many states, and the agent has noise-free data but a limited sampling frame, then his sample has value if and only if he knows the states have a common component that explains some of their prior variances.

This example highlights the importance of conceptual knowledge when making out-of-sample predictions. If the agent had no conceptual knowledge (and so assumed $\rho = 0$), then his extrapolation error could be arbitrarily large and the value of his sample could be arbitrarily small. Having

[50] We derive (A5)–(A7) in Appendix B.

conceptual knowledge allows him to use data on the on-support components of θ to learn about the off-support components. This lowers his extrapolation error and ensures his sample has *some* value, even if the sampling frame is limited.

A3 KL divergences

In Section 3.2, we claim the KL divergence (7) quantifies how much the agent's conceptual knowledge allows him to reduce θ. Proposition A4 justifies this claim. It says the KL divergence from the true prior $\mathbb{P}$ to the naïve prior $\mathbb{P}^{(0)}$ is (weakly) larger when states are more reducible (i.e., when the eigenvalues $\lambda_1, \ldots, \lambda_K$ of Σ are more spread out).

Proposition A4. *The KL divergence from* $\mathbb{P}$ *to* $\mathbb{P}^{(0)}$

(i) is non-negative,

(ii) equals zero when $\lambda_1, \ldots, \lambda_K$ *are equal, and*

(iii) does not fall when $\lambda_1, \ldots, \lambda_K$ *undergo a MPS.*

For example, suppose Σ is the matrix (8) constructed in Example 1. This matrix has eigenvalues $\lambda_1 = (1 + \rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$, which equal $\overline{\lambda} = \sigma^2$ when $\rho = 0$ and undergo a MPS when ρ rises (see Section 4.4). So, by Proposition A4, the KL divergence

$$\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = -\frac{\ln(1+\rho(K-1)) + (K-1)\ln(1-\rho)}{2}$$

from the true prior $\mathbb{P}$ to the naïve prior $\mathbb{P}^{(0)}$ must equal zero when $\rho = 0$ and be non-decreasing in ρ. Indeed

$$\begin{aligned}\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)})\Big|_{\rho=0} &= -\frac{\ln(1) + (K-1)\ln(1)}{2} \\ &= 0,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\rho}\mathcal{D}_{\text{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) &= \frac{K-1}{2}\left(\frac{1}{1-\rho} - \frac{1}{1+\rho(K-1)}\right) \\ &\geq 0\end{aligned}$$

with equality if and only if $\rho = 0$.

B Proofs

B1 Claims in Section 4

B1.1 Proof of Lemma 1

We have

$$\begin{aligned}\mathbb{E}[L(\theta, a') \mid \mathcal{S}] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[(\theta_k - a'_k)^2 \mid \mathcal{S}\right] \\ &= \frac{1}{K}\sum_{k=1}^{K} \Big(\big(\mathbb{E}[\theta_k \mid \mathcal{S}] - a'_k\big)^2 + \mathbb{V}(\theta_k \mid \mathcal{S}) \Big)\end{aligned}$$

for all $a' \in \mathbb{R}^K$. So $\mathbb{E}[L(\theta, a') \mid \mathcal{S}]$ attains its minimum value

$$\min_{a' \in \mathbb{R}^K} \mathbb{E}[L(\theta, a') \mid \mathcal{S}] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{V}(\theta_k \mid \mathcal{S})$$

when $a'_k = \mathbb{E}[\theta_k \mid \mathcal{S}]$ for each $k \in \{1, \ldots, K\}$. □

B1.2 Proof of Lemma 2

Our proof of Lemma 2 uses a well-known property of normally distributed random variables:

Lemma B1. *Let $n_1 \geq 1$ and $n_2 \geq 1$ be integers, and let $z \in \mathbb{R}^{n_1+n_2}$ be normally distributed with mean μ and variance Σ. Partition $z = (z_1, z_2)$ into vectors $z_1 \in \mathbb{R}^{n_1}$ and $z_2 \in \mathbb{R}^{n_2}$, and let $\mu = (\mu_1, \mu_2)$ and*

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

be the corresponding partitions of μ and Σ. If Σ_{22} is invertible, then

$$z_1 \mid z_2 \sim \mathcal{N}\Big(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(z_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big).$$

Proof. See Bishop (2006, p. 87) or DeGroot (2004, p. 55).

Proof of Lemma 2. Let $y \equiv (y^{(1)}, \ldots, y^{(n)})$ and $u \equiv (u^{(1)}, \ldots, u^{(n)})$ be the n-vectors of outcomes and errors, and let

$$W \equiv \begin{bmatrix} w^{(1)} & \cdots & w^{(n)} \end{bmatrix}^T$$

be the $n \times K$ design matrix. Then we can write (2) in vector form as

$$y \equiv W\theta + u.$$

Consider the concatenation of θ and y. It is normally distributed with variance

$$\mathbb{V}\left(\begin{bmatrix} \theta \\ y \end{bmatrix} \mid W\right) = \begin{bmatrix} \Sigma & \Sigma W^T \\ W\Sigma & W\Sigma W^T + \sigma_u^2 I_n \end{bmatrix}$$

under the agent's prior. Since observing $\mathcal{S}$ is equivalent to observing W and y, Lemma B1 implies

$$\begin{aligned}\mathbb{V}(\theta \mid \mathcal{S}) &= \mathbb{V}(\theta \mid W, y) \\ &= \Sigma - \Sigma W^T \left(W\Sigma W^T + \sigma_u^2 I_n\right)^{-1} W\Sigma \\ &= \left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right)^{-1}\end{aligned}$$

because $G = W^T W$. □

B1.3 Proof of Proposition 1

Our proof of Proposition 1 uses the following fact about sums of real, symmetric matrices.

Lemma B2. *Let $n \geq 1$ be an integer, let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be symmetric matrices with eigenvalues $a_1 \geq \cdots \geq a_n$ and $b_1 \geq \cdots \geq b_n$, and let $C = A + B$ have eigenvalues $c_1 \geq \cdots \geq c_n$. Then*

$$\sum_{j=1}^{k}(a_j + b_{n-j+1}) \leq \sum_{j=1}^{k} c_j \leq \sum_{j=1}^{k}(a_j + b_j)$$

for each $k \in \{1, \ldots, n\}$, with equality when $k = n$.

Proof. See Horn and Johnson (2012, Theorem 4.3.47).

Proof of Proposition 1. Now

$$\pi(\mathcal{S}) = \frac{1}{K}\left(\operatorname{tr}(\Sigma) - \operatorname{tr}\left(\left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right)^{-1}\right)\right)$$

by Lemma 2. Moreover, defining $Z \equiv V^T\Omega$ gives

$$\begin{aligned}\left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right)^{-1} &= \left(V\Lambda^{-1}V^T + \frac{1}{\sigma_u^2} VV^T\Omega\Delta\Omega^T VV^T\right)^{-1} \\ &= V\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z\Delta Z^T\right)^{-1} V^T\end{aligned}$$

and hence

$$\operatorname{tr}\left(\left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right)^{-1}\right) = \operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z\Delta Z^T\right)^{-1}\right)$$

by the orthogonality of V and the cyclic property of matrix traces. So (17) is equivalent to

$$\sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-1} \overset{\star\star}{\leq} \operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z\Delta Z^T\right)^{-1}\right) \overset{\star}{\leq} \sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2}\right)^{-1}. \tag{B1}$$

Now Λ^{-1} is real, symmetric, and positive definite. It has k^{th} largest eigenvalue $a_k \equiv 1/\lambda_{K-k+1} > 0$. Moreover, since Z is orthogonal, the matrix

$$B \equiv \frac{1}{\sigma_u^2} Z \Delta Z^T$$

is real, symmetric, and positive semi-definite. It has k^{th} largest eigenvalue $b_k \equiv \delta_k/\sigma_u^2 \geq 0$. Define

$$\begin{aligned} c_k^{**} &\equiv a_k + b_{K-k+1} \\ &= \frac{1}{\lambda_{K-k+1}} + \frac{\delta_{K-k+1}}{\sigma_u^2} \\ &> 0 \end{aligned}$$

and

$$\begin{aligned} c_k^{*} &\equiv a_k + b_k \\ &= \frac{1}{\lambda_{K-k+1}} + \frac{\delta_k}{\sigma_u^2} \\ &> 0 \end{aligned}$$

for each $k \in \{1, \ldots, K\}$, and consider the matrix $C \equiv \Lambda^{-1} + B$ with k^{th} largest eigenvalue c_k. This matrix is positive definite and so $c_k > 0$ for each k. Moreover, by Lemma B2, we have

$$\sum_{j=1}^{k} c_j^{**} \leq \sum_{j=1}^{k} c_j \leq \sum_{j=1}^{k} c_j^{*}$$

for each $k \in \{1, \ldots, K\}$, with equality when $k = K$.

Now define $g(z) \equiv 1/z$ for all $z > 0$. Then $g : (0, \infty) \to \mathbb{R}$ is convex. So, by Lemma 3, we have

$$\sum_{k=1}^{K} \frac{1}{c_k^{**}} \leq \sum_{k=1}^{K} \frac{1}{c_k} \leq \sum_{k=1}^{K} \frac{1}{c_k^{*}}. \tag{B2}$$

But

$$\sum_{k=1}^{K} \frac{1}{c_k^{**}} = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1}$$

and

$$\sum_{k=1}^{K} \frac{1}{c_k^{*}} = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2} \right)^{-1}$$

by the definitions of $c_1^{**}, \ldots, c_K^{**}$ and $c_1^{*}, \ldots, c_K^{*}$, and

$$\begin{aligned} \sum_{k=1}^{K} \frac{1}{c_k} &= \operatorname{tr}\left(C^{-1} \right) \\ &= \operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T \right)^{-1} \right) \end{aligned}$$

by the definition of C. Substituting these expressions into (B2) yields (B1), from which (17) follows.

It remains to show when the bounds $\star$ and $\star\star$ hold with equality.

Suppose $\omega_k = v_{K-k+1}$ for each $k \in \{1, \ldots, K\}$. Then

$$\begin{aligned} Z &= \begin{bmatrix} v_1 & \cdots & v_K \end{bmatrix}^T \begin{bmatrix} v_K & \cdots & v_1 \end{bmatrix} \\ &= \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix} \end{aligned}$$

is the $K \times K$ anti-diagonal matrix with jk^{th} entry

$$Z_{jk} = \begin{cases} 1 & \text{if } j + k = K + 1 \\ 0 & \text{if } j + k \neq K + 1. \end{cases}$$

So the inverse of

$$\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T = \Lambda^{-1} + \frac{1}{\sigma_u^2} \begin{bmatrix} \delta_K & & \\ & \ddots & \\ & & \delta_K \end{bmatrix}$$

has trace

$$\operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T \right)^{-1} \right) = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_{K-k+1}}{\sigma_u^2} \right)^{-1}$$

and thus $\star$ holds with equality.

Now suppose $\omega_k = v_k$ for each $k \in \{1, \ldots, K\}$. Then Z equals the $K \times K$ identity matrix. So the inverse of

$$\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T = \Lambda^{-1} + \frac{1}{\sigma_u^2} \Delta$$

has trace

$$\operatorname{tr}\left(\left(\Lambda^{-1} + \frac{1}{\sigma_u^2} Z \Delta Z^T \right)^{-1} \right) = \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1}$$

and thus $\star\star$ holds with equality. □

B1.4 Proof of Proposition 2

Consider the constrained minimization problem (18). We can ignore the constraint that δ_k is non-increasing in k because it does not bind (see below). So the problem has Lagrangian

$$\mathcal{L} \equiv \sum_{k=1}^{K} \left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2} \right)^{-1} - \sum_{k=1}^{K} \eta_k \delta_k - \eta \left(n - \sum_{k=1}^{K} \delta_k \right),$$

where $\eta_k \geq 0$ is the Lagrange multiplier on the non-negativity constraint $\delta_k \geq 0$ and $\eta \in \mathbb{R}$ is the multiplier on the sum constraint. Now

$$\frac{\partial^2 \mathcal{L}}{\partial \delta_j \, \partial \delta_k} = \begin{cases} \frac{2}{\sigma_u^4}\left(\frac{1}{\lambda_k} + \frac{\delta_k}{\sigma_u^2}\right)^{-3} & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

for each pair $(j,k) \in \{1,\ldots,K\}^2$, from which it follows that $\mathcal{L}$ is convex in the vector $(\delta_1,\ldots,\delta_K)$ whenever it has non-negative components. So if $\delta_1^*,\ldots,\delta_K^*$ solve (18), then they satisfy the first-order conditions (FOCs)

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \delta_k} \\ &= -\frac{1}{\sigma_u^2}\left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2}\right)^{-2} - \eta_k + \eta, \end{aligned}$$

the complementary slackness conditions $0 = \eta_k \delta_k^*$, and the sum constraint $\delta_1^* + \cdots + \delta_K^* = n$.

Suppose the non-negativity constraint on δ_k binds. Then the FOCs and complementary slackness conditions imply

$$\begin{aligned} 0 &< \eta_k \\ &= \eta - \frac{\lambda_k^2}{\sigma_u^2}, \end{aligned}$$

which holds if and only if $\lambda_k < \sigma_u\sqrt{\eta}$. But λ_k is non-increasing in k and the FOCs imply that η is strictly positive. So there is an integer $k_0 \in \{1,\ldots,K\}$ such that $\delta_k^* > 0$ if and only if $k \leq k_0$.

Suppose $k \leq k_0$. Then $\eta_k = 0$ and so the FOCs imply

$$\frac{\sigma_u^2}{\sqrt{\eta}} = \frac{\sigma_u^2}{\lambda_k} + \delta_k^*.$$

The left-hand side is constant in k, from which it follows that

$$\frac{\sigma_u^2}{\lambda_1} + \delta_1^* = \frac{\sigma_u^2}{\lambda_k} + \delta_k^*$$

and therefore

$$\delta_k^* = \delta_1^* + \sigma_u^2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right).$$

Then the sum constraint implies

$$\begin{aligned} n &= \sum_{k=1}^{k_0}\left(\delta_1^* + \sigma_u^2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right)\right) \\ &= k_0\delta_1^* + \sigma_u^2 \sum_{k=1}^{k_0}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right). \end{aligned}$$

Thus

$$\begin{aligned}\delta_k^* &= \frac{1}{k_0}\left(n - \sigma_u^2 \sum_{j=1}^{k_0}\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_j}\right)\right) + \sigma_u^2\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_k}\right) \\ &= \frac{n}{k_0} + \sigma_u^2\left(\frac{1}{k_0}\sum_{j=1}^{k_0}\frac{1}{\lambda_j} - \frac{1}{\lambda_k}\right)\end{aligned}$$

for each $k \leq k_0$ and $\delta_k^* = 0$ for each $k > k_0$. Then

$$\begin{aligned}\sum_{k=1}^{K}\left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2}\right)^{-1} &= \sum_{k=1}^{k_0}\left(\frac{1}{\lambda_k} + \frac{1}{\sigma_u^2}\left(\frac{n}{k_0} + \sigma_u^2\left(\frac{1}{k_0}\sum_{j=1}^{k_0}\frac{1}{\lambda_j} - \frac{1}{\lambda_k}\right)\right)\right)^{-1} + \sum_{k>k_0}\left(\frac{1}{\lambda_k} + 0\right)^{-1} \\ &= k_0^2\left(\sum_{j=1}^{k_0}\frac{1}{\lambda_j} + \frac{n}{\sigma_u^2}\right)^{-1} + \sum_{k>k_0}\lambda_k\end{aligned}$$

is non-increasing in k_0 when $k_0 \leq R^*$. Thus, the eigenvalues $\delta_1^*, \ldots, \delta_K^*$ defined by (21) solve (18). They are non-increasing because $\lambda_1, \ldots, \lambda_K$ are non-increasing. Moreover, Proposition 1 implies

$$\begin{aligned}\pi(\mathcal{S}) &\leq \frac{1}{K}\sum_{k=1}^{K}\left(\lambda_k - \left(\frac{1}{\lambda_k} + \frac{\delta_k^*}{\sigma_u^2}\right)^{-1}\right) \\ &= \frac{1}{K}\left(\sum_{k=1}^{K}\lambda_k - \left((R^*)^2\left(\sum_{j=1}^{R^*}\frac{1}{\lambda_j} + \frac{n}{\sigma_u^2}\right)^{-1} + \sum_{k>R^*}\lambda_k\right)\right) \\ &= \pi^*,\end{aligned}$$

with equality if (22) holds. □

B1.5 Proof of Lemma 3

The result follows from establishing three equivalences:

1. *(i)* $\iff$ *(ii)*. Rothschild and Stiglitz (1970, Theorem 2) show that (i) is equivalent to

   (ii′) $\int_0^\infty g(z)\,\mathrm{d}F'(z) \geq \int_0^\infty g(z)\,\mathrm{d}F(z)$ for all convex functions $g : (0,\infty) \to \mathbb{R}$,

   which is equivalent to (ii) by the definitions of F and F'.

2. *(ii)* $\iff$ *(iii)*. Consider the K-vectors $\lambda' \equiv (\lambda_1', \ldots, \lambda_K')$ and $\lambda \equiv (\lambda_1, \ldots, \lambda_K)$. Arnold (1987, Theorem 2.9) shows that (ii) holds precisely when λ' majorizes λ. But the components of λ' and λ are non-increasing, and so λ' majorizes λ if and only if (iii) holds.

3. *(iii)* $\iff$ *(iv)*. For each $k \in \{1, \ldots, K\}$ we have

$$\begin{aligned}\sum_{j=1}^{k}\lambda_j' - \sum_{j=1}^{k}\lambda_j &= \left(\sum_{j=1}^{K}\lambda_j' - \sum_{j>k}\lambda_j'\right) - \left(\sum_{j=1}^{K}\lambda_j - \sum_{j>k}\lambda_j\right) \\ &= \left(\sum_{j=1}^{K}\lambda_j' - \sum_{j=1}^{K}\lambda_j\right) - \left(\sum_{j>k}\lambda_j' - \sum_{j>k}\lambda_j\right),\end{aligned}$$

from which it follows that (iii) and (iv) are equivalent. □

B2 Claims in Sections 5–7

B2.1 Proof of Theorem 1

Our proof of Theorem 1 invokes the following lemma.

Lemma B3. *The value π^* of an optimal sample does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS.*

Proof. Now

$$\begin{aligned} R^* &\in \operatorname*{arg\,min}_{k_0 \in \{1,\ldots,K\}} \left(k_0 \left(\frac{1}{k_0} \left(\sum_{k=1}^{k_0} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right) \right)^{-1} + \sum_{k>k_0} \lambda_k \right) \\ &= \operatorname*{arg\,max}_{k_0 \in \{1,\ldots,K\}} \left(\sum_{k=1}^{k_0} \lambda_k - k_0^2 \left(\sum_{k=1}^{k_0} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} \right) \end{aligned}$$

from the proof of Proposition 2. So if $\lambda_1, \ldots, \lambda_K$ undergo a MPS, then R^* changes only if doing so makes $\mathcal{S}$ more valuable. So it suffices to show that for fixed R^*, the MPS does not lower the RHS of (20).

Let $\lambda_1' \geq \cdots \geq \lambda_K' > 0$ be the eigenvalues after the MPS. By Lemma 3, the difference

$$\eta \equiv \sum_{k=1}^{R^*} \lambda_k' - \sum_{k=1}^{R^*} \lambda_k \tag{B3}$$

is non-negative. The MPS raises the first bracketed term on the RHS of (20) by η. So it suffices to show that the MPS lowers the second bracketed term by at most η:

$$\underbrace{(R^*)^2 \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k'} + \frac{n}{\sigma_u^2} \right)^{-1}}_{S'} - \underbrace{(R^*)^2 \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1}}_{S} \leq \eta. \tag{B4}$$

Consider the first term S' on the LHS of (B4). This term is largest when the harmonic sum

$$H' \equiv \sum_{k=1}^{R^*} \frac{1}{\lambda_k'}$$

is smallest. Defining $\eta_k \equiv \lambda_k' - \lambda_k$ for each $k \in \{1, \ldots, K\}$ gives

$$H' \equiv \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k}$$

and $\eta_1 + \cdots + \eta_{R^*} = \eta$. Lemma 3 implies

$$\sum_{j=1}^{k} \eta_j \geq 0$$

for each $k \in \{1, \ldots, R^*\}$. Thus

$$H' \geq H^* \equiv \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k^*},$$

where $\eta_1^*, \ldots, \eta_{R^*}^*$ solve the constrained minimization problem

$$\begin{aligned} \min_{\eta_1, \ldots, \eta_{R^*} \in \mathbb{R}} & \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k} \\ \text{subject to} \quad & \lambda_k + \eta_k > 0 \text{ for each } k \in \{1, \ldots, R^*\}, \\ & \sum_{j=1}^{k} \eta_j \geq 0 \text{ for each } k \in \{1, \ldots, R^*\}, \\ \text{and} \quad & \sum_{k=1}^{R^*} \eta_k = \eta. \end{aligned} \tag{B5}$$

Setting $\lambda_k' = \lambda_k + \eta_k^*$ for each $k \in \{1, \ldots, R^*\}$ yields the "worst-case" MPS that maximizes the first term S' on the LHS of (B4) given the difference (B3).

The differences $\eta_1^*, \ldots, \eta_{R^*}^*$ that solve (B5) are non-negative. To see why, notice that $\eta_1 < 0$ is infeasible and assume towards a contradiction that $\eta_\ell^* < 0 \leq \min\{\eta_1^*, \ldots, \eta_{\ell-1}^*\}$ for some $\ell > 1$. Then

$$\ell' \equiv \max\{k \in \{1, \ldots, \ell - 1\} : \eta_k^* > 0\}$$

must exist, for otherwise $\eta_1^*, \ldots, \eta_{R^*}^*$ would violate the constraint

$$\sum_{j=1}^{\ell} \eta_k^* \geq 0.$$

Defining

$$\eta_k^\dagger \equiv \begin{cases} \eta_{\ell'}^* + \eta_\ell^* & \text{if } k = \ell' \\ 0 & \text{if } \ell' < k = \ell \\ \eta_k^* & \text{otherwise} \end{cases}$$

gives

$$\sum_{j=1}^{k} \eta_j^\dagger = \begin{cases} \sum_{j=1}^{\ell} \eta_j^* & \text{if } \ell' \leq k \leq \ell \\ \sum_{j=1}^{k} \eta_j^* & \text{otherwise} \end{cases}$$

for each $k \in \{1, \ldots, R^*\}$, from which it follows that $\eta_1^\dagger, \ldots, \eta_{R^*}^\dagger$ are feasible. But $\lambda_{\ell'} \geq \lambda_\ell$ and $\eta_{\ell'}^* > 0$, and so $\lambda_{\ell'} + \eta_{\ell'}^* > \lambda_\ell > 0$. Thus

$$\frac{1}{\lambda_{\ell'} + \eta_{\ell'}^* + \eta_\ell^*} + \frac{1}{\lambda_\ell} < \frac{1}{\lambda_{\ell'} + \eta_{\ell'}^*} + \frac{1}{\lambda_\ell + \eta_\ell^*}$$

because $g(z) \equiv 1/z$ is a strictly decreasing and convex function of $z > 0$. But then

$$\begin{aligned}\sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k^\dagger} &= \sum_{k<\ell'} \frac{1}{\lambda_k + \eta_k^*} + \frac{1}{\lambda_{\ell'} + \eta_{\ell'}^* + \eta_\ell^*} + \frac{1}{\lambda_\ell} + \sum_{k>\ell} \frac{1}{\lambda_k + \eta_k^*} \\ &< \frac{1}{\lambda_{\ell'} + \eta_{\ell'}^*} + \frac{1}{\lambda_\ell + \eta_\ell^*} + \sum_{k \notin \{\ell', \ell\}} \frac{1}{\lambda_k + \eta_k^*} \\ &= \sum_{k=1}^{R^*} \frac{1}{\lambda_k + \eta_k^*},\end{aligned}$$

contradicting the optimality of $\eta_1^*, \ldots, \eta_{R^*}^*$. So they must be non-negative because ℓ cannot exist.

Finally, we use the non-negativity of $\eta_1^*, \ldots, \eta_{R^*}^*$ to establish the upper bound (B4) on $(S' - S)$. Let $k \in \{1, \ldots, R^*\}$ and consider the derivative

$$\frac{\partial S}{\partial \lambda_k} = \left(\frac{R^*}{\lambda_k} \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \right)^{-1} \right)^2$$

of S with respect to λ_k. This derivative is non-negative. It is also bounded above by one, since

$$\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2} \geq \frac{R^*}{\lambda_k}$$

by the definition of R^*. So S is a 1-Lipschitz function of $\lambda_1, \ldots, \lambda_{R^*}$: changing λ_k by η_k changes S by at most $|\eta_k|$. Letting S^* be the value of S that obtains from changing λ_k by η_k^* gives

$$\begin{aligned}S' - S &\overset{\star}{\leq} S^* - S \\ &\leq |S^* - S| \\ &\overset{\star\star}{\leq} \sum_{k=1}^{K} |\eta_k^*|,\end{aligned}$$

where $\star$ uses the maximality of S^* (induced by the minimality of H^*) and $\star\star$ uses the Lipschitz property. But $\eta_1^*, \ldots, \eta_{R^*}^*$ are non-negative and sum to η, from which the bound (B4) follows:

$$\begin{aligned}S' - S &\leq \sum_{k=1}^{K} \eta_k^* \\ &= \eta.\end{aligned}$$

□

Proof of Theorem 1. It suffices to prove (ii) and (iii), which together imply (i). This is because every distribution of $\lambda_1, \ldots, \lambda_K$ is a MPS of the degenerate distribution under which they are equal (to their mean $\overline{\lambda}$).

Consider (ii). If $\lambda_1, \ldots, \lambda_K$ are equal, then $\lambda_k = \overline{\lambda}$ for each $k \in \{1, \ldots, K\}$, and so $R^* = R^{(0)}$ and $\pi^* = \pi^{(0)}$ by definition. Thus $\Pi \equiv \pi^* - \pi^{(0)} = 0$.

Now consider (iii). The value $\pi^{(0)}$ of the naïve agent's optimal sample depends on $\lambda_1, \ldots, \lambda_K$ via their mean $\overline{\lambda}$ only. It does not change when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. Since π^* does not fall under the MPS (by Lemma B3), neither does $\Pi \equiv \pi^* - \pi^{(0)}$. □

B2.2 Proof of Theorem 2

Define

$$\tau_k \equiv \overline{\lambda}\left(\frac{k}{\lambda_k} - \sum_{j=1}^{k}\frac{1}{\lambda_j}\right)$$

for each $k \in \{1,\ldots,K\}$. Then $\tau_1 = 0$, and for each $k < K$ the difference

$$\tau_{k+1} - \tau_k = k\overline{\lambda}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right)$$

is non-negative because $\overline{\lambda} > 0$ and $\lambda_{k+1} \leq \lambda_k$. So τ_k is non-decreasing in k and hence

$$R^* = \max\{k \in \{1,\ldots,K\} : \tau_k \leq \tau\}$$

is non-decreasing in τ. Now define $\tau_{K+1} \equiv \infty$ and suppose $\tau \in [\tau_k, \tau_{k+1})$ for some $k \in \{1,\ldots,K\}$. Then $R^* = k$ and so

$$\begin{aligned}\Pi &= \Pi_k \\ &\equiv \frac{\overline{\lambda}}{K}\left(\sum_{j=1}^{k}\frac{\lambda_j}{\overline{\lambda}} - k^2\left(\sum_{j=1}^{k}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - \frac{K\tau}{K+\tau}\right).\end{aligned}$$

Each piece Π_k is continuous in τ. Moreover, for each $k < K$ the difference

$$\Pi_{k+1} - \Pi_k = -\frac{\overline{\lambda}}{K}\left((k+1)^2\left(\sum_{j=1}^{k+1}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - k^2\left(\sum_{j=1}^{k}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - \frac{\lambda_{k+1}}{\overline{\lambda}}\right)$$

between consecutive pieces converges to zero as $\tau \to \tau_{k+1}$. It follows that Π is continuous in τ. So to determine whether Π is increasing or decreasing in τ, it suffices to analyze its derivative

$$\frac{\partial \Pi_k}{\partial \tau} = \frac{\overline{\lambda}}{K}\left(k^2\left(\sum_{j=1}^{k}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-2} - \left(\frac{K}{K+\tau}\right)^2\right) \tag{B6}$$

on each piece Π_k.

Consider the final piece

$$\begin{aligned}\Pi_K &= \frac{\overline{\lambda}}{K}\left(\sum_{j=1}^{K}\frac{\lambda_j}{\overline{\lambda}} - K^2\left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1} - \frac{K\tau}{K+\tau}\right) \\ &= K\overline{\lambda}\left(\frac{1}{K+\tau} - \left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1}\right).\end{aligned}$$

If $\lambda_1, \ldots, \lambda_K$ are equal (i.e., if $\lambda_1 = \lambda_K$), then $\lambda_k = \overline{\lambda}$ and $\tau_k = 0$ for each $k \in \{1, \ldots, K\}$, and so

$$\begin{aligned}\Pi\Big|_{\lambda_1=\lambda_K} &= \Pi_K\Big|_{\lambda_1=\lambda_K} \\ &= K\overline{\lambda}\left(\frac{1}{K+\tau} - \left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\overline{\lambda}} + \tau\right)^{-1}\right) \\ &= 0\end{aligned}$$

for all $\tau \geq 0$. Whereas if $\lambda_1, \ldots, \lambda_K$ are not equal (i.e., if $\lambda_1 > \lambda_K$), then

$$\begin{aligned}\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j} &> \frac{K\overline{\lambda}}{\frac{1}{K}\sum_{j=1}^{K}\lambda_j} \\ &= K\end{aligned}$$

by Jensen's inequality and the definition of $\overline{\lambda}$, from which it follows that

$$\frac{\partial \Pi_K}{\partial \tau}\bigg|_{\lambda_1>\lambda_K} = K\overline{\lambda}\left(\left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-2} - \left(\frac{1}{K+\tau}\right)^2\right)$$

is strictly negative. Thus Π is non-increasing in τ whenever $\tau \geq \tau_K$. Moreover,

$$\begin{aligned}\lim_{\tau\to\infty}\Pi &= \lim_{\tau\to\infty}\Pi_K \\ &= K\overline{\lambda}\lim_{\tau\to\infty}\left(\frac{1}{K+\tau} - \left(\sum_{j=1}^{K}\frac{\overline{\lambda}}{\lambda_j} + \tau\right)^{-1}\right) \\ &= 0.\end{aligned}$$

So if $\lambda_1, \ldots, \lambda_K$ are equal, then $\tau_K = 0$ and the result follows from letting $\tau' = 0$.

It remains to show that if $\lambda_1, \ldots, \lambda_K$ are *not* equal, then there exists $\tau' \in (0, \tau_K)$ such that Π is increasing in τ if and only if $\tau < \tau'$.

Suppose $\tau \in [\tau_k, \tau_{k+1})$ for some $k < K$. Then Π is increasing in τ if and only if (B6) exceeds zero, which happens precisely when

$$\begin{aligned}\tau &< \tau_k' \\ &\equiv \frac{K}{K-k}\left(k\left(1 - \frac{\overline{\lambda}}{\lambda_k}\right) + \tau_k\right).\end{aligned}$$

So Π_k is decreasing in $\tau \in [\tau_k, \tau_{k+1})$ if $\tau_k' < \tau_k$, increasing if $\tau_k' \geq \tau_{k+1}$, and increasing-and-then-decreasing if $\tau_k \leq \tau_k' < \tau_{k+1}$. Now $\tau_k' \geq \tau_k$ if and only if

$$\frac{K-k}{\lambda_k} + \sum_{j=1}^{k}\frac{1}{\lambda_j} \leq \frac{K}{\overline{\lambda}},$$

whereas $\tau'_k < \tau_{k+1}$ if and only if

$$\frac{K}{\overline{\lambda}} < \frac{K-(k+1)}{\lambda_{k+1}} + \sum_{j=1}^{k+1} \frac{1}{\lambda_j}.$$

So defining

$$\eta_k \equiv \frac{K-k}{\lambda_k} + \sum_{j=1}^{k} \frac{1}{\lambda_j}$$

for each $k \in \{1, \ldots, K\}$ gives $\tau'_k \in [\tau_k, \tau_{k+1})$ if and only if $K/\overline{\lambda} \in [\eta_k, \eta_{k+1})$. But η_k is non-decreasing in k because $\lambda_{k+1} \leq \lambda_k$ and therefore

$$\begin{aligned} \eta_{k+1} - \eta_k &= (K-k)\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right) \\ &\geq 0. \end{aligned}$$

It follows that $\tau'_k \in [\tau_k, \tau_{k+1})$ for at most one $k < K$. But there is at least one such k when $\lambda_1, \ldots, \lambda_K$ are not equal. To see why, notice that

$$\begin{aligned} \lim_{\tau \to 0} \frac{\partial \Pi}{\partial \tau} &= \lim_{\tau \to 0} \frac{\partial \Pi_1}{\partial \tau} \\ &= \frac{\overline{\lambda}}{K} \lim_{\tau \to 0} \left(\left(\frac{\overline{\lambda}}{\lambda_1} + \tau \right)^{-2} - \left(\frac{K}{K+\tau} \right)^2 \right) \\ &= \frac{\overline{\lambda}}{K} \left(\left(\frac{\lambda_1}{\overline{\lambda}} \right)^2 - 1 \right) \end{aligned}$$

is strictly positive when $\lambda_1 > \overline{\lambda}$, which holds precisely when $\lambda_1, \ldots, \lambda_K$ are not equal, in which case the value Π is decreasing in τ whenever $\tau > \tau_K$. So Π is initially increasing in τ and eventually decreasing in τ, which, by continuity, means its derivative with respect to τ changes sign at least once. Therefore, if $\lambda_1, \ldots, \lambda_K$ are not equal, then there is a unique $k < K$ such that $\tau'_k \in [\tau_k, \tau_{k+1})$. Letting $\tau' = \tau'_k > 0$ completes the proof. □

B2.3 Proof of Proposition 3

Our proof of Proposition 3 invokes the following lemma.

Lemma B4. *Suppose θ has prior variance* (8) *with $\sigma^2 > 0$ and $\rho \in [0, 1)$.*

(i) There is a threshold $\rho' \in (0, 1)$ such that

$$R^* = \begin{cases} K & \text{if } \rho \leq \rho' \\ 1 & \text{if } \rho > \rho'. \end{cases} \tag{B7}$$

(ii) The value π^* of an optimal sample rises when ρ rises.

Proof. Consider (i). If $\lambda_1 \geq \lambda_2 = \cdots = \lambda_K$, then

$$R^* = \begin{cases} 1 & \text{if } \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} < \frac{1}{\lambda_2} \\ K & \text{otherwise.} \end{cases}$$

Now (8) has eigenvalues $\lambda_1 = (1+\rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$. So $R^* = K$ if and only if

$$\begin{aligned} 0 &\leq \frac{1}{(1+\rho(K-1))\sigma^2} - \frac{1}{(1-\rho)\sigma^2} + \frac{n}{\sigma_u^2} \\ &= \frac{1}{\sigma^2}\left(\frac{1}{1+\rho(K-1)} - \frac{1}{1-\rho} + \tau\right). \end{aligned}$$

The bracketed term on the RHS is continuous and decreasing in ρ, strictly positive when $\rho = 0$, and unbounded below as $\rho \to 1$. So, by the intermediate value theorem, there exists $\rho' \in (0,1)$ such that (B7) holds.

Now consider (ii). Substituting (B7) into (20) gives

$$\begin{aligned} \pi^* &= \frac{1}{K}\begin{cases} \sum_{k=1}^K \lambda_k - K^2\left(\sum_{k=1}^K \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1} & \text{if } \rho \leq \rho' \\ \lambda_1 - \left(\frac{1}{\lambda_1} + \frac{n}{\sigma_u^2}\right)^{-1} & \text{if } \rho > \rho' \end{cases} \\ &= \frac{\sigma^2}{K}\begin{cases} K - K^2\left(\frac{1}{1+\rho(K-1)} + \frac{K-1}{1-\rho} + \tau\right)^{-1} & \text{if } \rho \leq \rho' \\ 1+\rho(K-1) - \left(\frac{1}{1+\rho(K-1)} + \tau\right)^{-1} & \text{if } \rho > \rho', \end{cases} \end{aligned} \tag{B8}$$

which is piecewise increasing in ρ:

$$\begin{aligned} \frac{\partial}{\partial\rho}\left[\pi^*\big|_{\rho\leq\rho'}\right] &= K(K-1)\left(\frac{1}{1+\rho(K-1)} + \frac{K-1}{1-\rho} + \tau\right)^{-2}\left(\frac{1}{(1-\rho)^2} - \frac{1}{(1+\rho(K-1))^2}\right) \\ &\geq 0 \end{aligned}$$

with equality if and only if $\rho = 0$, and

$$\begin{aligned} \frac{\partial}{\partial\rho}\left[\pi^*\big|_{\rho>\rho'}\right] &= \frac{(K-1)\sigma^2}{K}\left(1 + \left(\frac{1}{1+\tau(1+\rho(K-1))}\right)^2\right) \\ &> 0. \end{aligned}$$

□

Proof of Proposition 3. Suppose the sample $\mathcal{S}$ is optimal. Then its value π^* equals

$$\pi^{(0)} = \frac{\sigma^2\tau}{K+\tau}$$

when $\rho = 0$. Now π^* is increasing in ρ (by Lemma B4), whereas $\pi^{(0)}$ is constant in ρ. So $\Pi = \pi^* - \pi^{(0)}$ equals zero when $\rho = 0$ and is increasing in ρ.

It remains to prove (iii). Now (8) has eigenvalues $\lambda_1 = (1+\rho(K-1))\sigma^2$ and $\lambda_2 = \cdots = \lambda_K = (1-\rho)\sigma^2$, which have mean $\overline{\lambda} = \sigma^2$. Defining

$$\begin{aligned}\tau_K &\equiv \overline{\lambda}\left(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}\right) \\ &= \frac{\rho K}{(1-\rho)(1+\rho(K-1))}\end{aligned}$$

gives

$$\begin{aligned}R^* &= \begin{cases} 1 & \text{if } \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} < \frac{1}{\lambda_K} \\ K & \text{if } \frac{1}{\lambda_1} + \frac{n}{\sigma_u^2} \geq \frac{1}{\lambda_K} \end{cases} \\ &= \begin{cases} 1 & \text{if } \tau < \tau_K \\ K & \text{if } \tau \geq \tau_K, \end{cases}\end{aligned}$$

which when substituted into (20) gives

$$\Pi = \frac{\sigma^2}{K}\begin{cases} 1+\rho(K-1) - \left(\frac{1}{1+\rho(K-1)} + \tau\right)^{-1} - \frac{K\tau}{K+\tau} & \text{if } \tau < \tau_K \\ K^2\left((K+\tau)^{-1} - \left(\sum_{k=1}^K \frac{\sigma^2}{\lambda_k} + \tau\right)^{-1}\right) & \text{if } \tau \geq \tau_K. \end{cases}$$

The first piece is (weakly) concave in τ: differentiating it with respect to τ gives

$$\frac{\partial}{\partial \tau}\left[\Pi\big|_{\tau<\tau_K}\right] = \frac{\sigma^2}{K}\left(\left(\frac{1}{1+\rho(K-1)} + \tau\right)^{-2} - \left(\frac{K}{K+\tau}\right)^2\right),$$

which is strictly positive if and only if

$$\tau < \tau' \equiv \frac{\rho K}{1+\rho(K-1)}.$$

In contrast, our proof of Theorem 2 shows that the second piece (with $\tau \geq \tau_K$) is non-increasing in τ. But $\tau' \leq \tau_K$, from which (iii) follows. □

B2.4 Proof of Lemma 4

Define

$$\tau_k \equiv \overline{\lambda}\left(\frac{k}{\lambda_k} - \sum_{j=1}^k \frac{1}{\lambda_j}\right)$$

for each $k \in \{1, \ldots, K\}$ so that

$$R^* = \max\{k \in \{1, \ldots, K\} : \tau_k \leq \tau\}$$

as in the proof of Theorem 2. Fix $J \in \{0, \ldots, K\}$ and define

$$\begin{aligned}\tau_k^{(J)} &\equiv \overline{\lambda}\left(\frac{k}{\lambda_k^{(J)}} - \sum_{j=1}^{k} \frac{1}{\lambda_j^{(J)}}\right) \\ &= \overline{\lambda}\begin{cases} 0 & \text{if } J = 0 \\ \frac{k}{\lambda_k} - \sum_{j=1}^{k} \frac{1}{\lambda_j} & \text{if } J > 0 \text{ and } k \leq J \\ \frac{J}{\lambda_K^{(J)}} - \sum_{j=1}^{J} \frac{1}{\lambda_j} & \text{if } J > 0 \text{ and } k > J. \end{cases}\end{aligned}$$

for each $k \in \{1, \ldots, K\}$. Then $\tau_1^{(J)} = 0$, and for each $k < K$ the difference

$$\tau_{k+1}^{(J)} - \tau_k^{(J)} = \overline{\lambda}\begin{cases} 0 & \text{if } J = 0 \\ k\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right) & \text{if } J > 0 \text{ and } k \leq J - 1 \\ J\left(\frac{1}{\lambda_K^{(J)}} - \frac{1}{\lambda_J}\right) & \text{if } J > 0 \text{ and } k = J \\ 0 & \text{if } J > 0 \text{ and } k > J. \end{cases}$$

is non-negative because $\lambda_{k+1} \leq \lambda_k$ and $\lambda_K^{(J)} \leq \lambda_J$. So $\tau_k^{(J)}$ is non-decreasing in k and

$$R^{(J)} = \max\left\{k \in \{1, \ldots, K\} : \tau_k^{(J)} \leq \tau\right\}.$$

Define $\tau_0^{(0)} \equiv 0$ and notice $\tau_J^{(J)} = \cdots = \tau_K^{(J)}$. So if $\tau \geq \tau_J^{(J)}$, then $R^{(J)} = K$; if $\tau < \tau_J^{(J)}$, then

$$\begin{aligned} R^{(J)} &= \max\left\{k \in \{1, \ldots, J\} : \tau_k^{(J)} \leq \tau\right\} \\ &= \min\{J, \max\{k \in \{1, \ldots, K\} : \tau_k \leq \tau\}\} \\ &= \min\{J, R^*\}. \end{aligned}$$

But if $J < K$, then

$$\begin{aligned}\tau_{J+1}^{(J+1)} - \tau_J^{(J)} &= \overline{\lambda}\left(\left(\frac{J+1}{\lambda_{J+1}} - \sum_{j=1}^{J+1} \frac{1}{\lambda_j}\right) - \left(\frac{J}{\lambda_J} - \sum_{j=1}^{J} \frac{1}{\lambda_j}\right)\right) \\ &= J\overline{\lambda}\left(\frac{1}{\lambda_{J+1}} - \frac{1}{\lambda_J}\right)\end{aligned}$$

is non-negative because $\lambda_{J+1} \leq \lambda_J$. So $\tau_J^{(J)}$ is non-decreasing in J, from which it follows that

$$J' \equiv \max\left\{k \in \{0, \ldots, K\} : \tau_j^{(j)} \leq \tau \text{ for each } j \in \{0, \ldots, k\}\right\}$$

exists and

$$R^{(J)} = \begin{cases} K & \text{if } J \leq J' \\ \min\{J, R^*\} & \text{if } J > J'. \end{cases}$$

Clearly J' is non-decreasing in τ. □

B2.5 Proof of Theorem 3

Now (i) follows from (ii)–(iv), while (ii) follows from the definition of $\Pi^{(J)} = \pi^{(J)} - \pi^{(0)}$ and (iii) follows from Lemma 5(i). For (iv), suppose $J \geq R^*$. Then $R^{(J)} = R^*$ by Lemma 4, so

$$\begin{aligned}
\pi^{(J)} &= \frac{1}{K}\left(\sum_{k=1}^{R^*} \lambda_k^{(J)} - (R^*)^2 \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \\
&= \frac{1}{K}\left(\sum_{k=1}^{R^*} \lambda_k - (R^*)^2 \left(\sum_{k=1}^{R^*} \frac{1}{\lambda_k} + \frac{n}{\sigma_u^2}\right)^{-1}\right) \\
&= \pi^*
\end{aligned}$$

and hence $\Pi^{(J)} = \Pi$ by definition. □

B2.6 Proof of Lemma 5

We prove (i) and (ii) separately:

(i) It suffices to show that $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS when J rises. Then (i) follows from an argument similar to that used to prove Lemma B3.

Fix $J < K$. For each $k \in \{1, \ldots, K\}$ we have

$$\lambda_k^{(J+1)} - \lambda_k^{(J)} = \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ \lambda_K^{(J+1)} - \lambda_K^{(J)} & \text{if } k > J+1 \end{cases}$$

and hence

$$\begin{aligned}
\sum_{j=1}^{k} \lambda_j^{(J+1)} - \sum_{j=1}^{k} \lambda_j^{(J)} &= \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ \lambda_{J+1} - \lambda_K^{(J)} + (k-(J+1))\left(\lambda_K^{(J+1)} - \lambda_K^{(J)}\right) & \text{if } k > J+1 \end{cases} \\
&= \begin{cases} 0 & \text{if } k \leq J \\ \lambda_{J+1} - \lambda_K^{(J)} & \text{if } k = J+1 \\ (K-k)\left(\lambda_K^{(J)} - \lambda_K^{(J+1)}\right) & \text{if } k > J+1. \end{cases} \qquad \text{(B9)}
\end{aligned}$$

because $\lambda_{J+1} = (K-J)\lambda_K^{(J)} - (K-(J+1))\lambda_K^{(J+1)}$. We also have

$$\begin{aligned}
\lambda_{J+1} &= \frac{1}{K-J}\sum_{k>J} \lambda_{J+1} \\
&\geq \frac{1}{K-J}\sum_{k>J} \lambda_k \\
&= \lambda_K^{(J)}
\end{aligned}$$

because $\lambda_{J+1} \geq \cdots \geq \lambda_K$. Likewise $\lambda_{J+1} \geq \lambda_K^{(J+1)}$ and so

$$\begin{aligned}
\lambda_K^{(J)} - \lambda_K^{(J+1)} &= \frac{1}{K-J}\sum_{k>J}\lambda_k - \frac{1}{K-(J+1)}\sum_{k>J+1}\lambda_k \\
&= \frac{1}{K-J}\lambda_{J+1} + \left(\frac{1}{K-J} - \frac{1}{K-(J+1)}\right)\sum_{k>J+1}\lambda_k \\
&= \frac{1}{K-J}\lambda_{J+1} - \frac{1}{K-J}\lambda_K^{(J+1)} \\
&\geq 0.
\end{aligned}$$

So the difference (B9) is non-negative and equals zero when $k = K$. Thus, by Lemma 3, the eigenvalues $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS when J rises.

(ii) Fix $J \in \{0, \ldots, K\}$ and define

$$t_k^{(J)} \equiv k^2\left(\sum_{j=1}^{k}\frac{1}{\lambda_j^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-1} + \sum_{j>k}\lambda_j^{(J)}$$

for each $k \in \{1, \ldots, K\}$. Then

$$\pi^{(J)} = \overline{\lambda} - \frac{1}{K}\min\left\{t_k^{(J)} : k \in \{1, \ldots, K\}\right\}$$

from the proof of Proposition 2. But

$$\frac{\partial t_k^{(J)}}{\partial n} = -\frac{k^2}{\sigma_u^2}\left(\sum_{j=1}^{k}\frac{1}{\lambda_j^{(J)}} + \frac{n}{\sigma_u^2}\right)^{-2}$$

is strictly negative, from which it follows that $\pi^{(J)}$ is increasing in n. □

B2.7 Proof of Theorem 4

Our proof of Theorem 4 invokes the following Lemma:

Lemma B5. *Fix $J \in \{0, \ldots, K\}$. Then $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS when $\lambda_1, \ldots, \lambda_K$ undergo a MPS.*

Proof. Fix $k \in \{1, \ldots, K\}$. By Lemma 3, the cumulative sum

$$\sum_{j=1}^{\min\{k,J\}}\lambda_j$$

does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS, while the tail sum

$$\sum_{j>J}\lambda_j$$

does not rise under the MPS. So the MPS does not lower

$$\sum_{j=1}^{k} \lambda_j^{(J)} = \begin{cases} \sum_{j=1}^{k} \lambda_j & \text{if } k \leq J \\ \sum_{j=1}^{J} \lambda_j + (k-J)\lambda_K^{(J)} & \text{if } k > J \end{cases}$$
$$= \begin{cases} \sum_{j=1}^{\min\{k,J\}} \lambda_j & \text{if } k \leq J \\ \sum_{j=1}^{K} \lambda_j - \frac{K-k}{K-J} \sum_{j>J} \lambda_j & \text{if } k > J \end{cases}$$

and leaves it unchanged when $k = K$. The result follows from Lemma 3. □

Proof of Theorem 4. We prove (i) and (ii) separately:

(i) Fix $n \geq 0$. Now $\pi^{(J)}$ is non-decreasing in J (by Lemma 5), so if $\pi^{(J)} \geq \pi_0$ then $\pi^{(J+1)} \geq \pi_0$. Thus

$$\{n \geq 0 : \pi^{(J)} \geq \pi_0\} \subseteq \{n \geq 0 : \pi^{(J+1)} \geq \pi_0\}$$

and therefore $n_{\pi_0}^{(J)} \geq n_{\pi_0}^{(J+1)}$.

(ii) It suffices to show that $\pi^{(J)}$ does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. Then, since $\pi^{(J)}$ is increasing in n (by Lemma 5), the MPS expands $\{n \geq 0 : \pi^{(J)} \geq \pi_0\}$ and so cannot raise $n_{\pi_0}^{(J)}$. But the argument used to prove Lemma B3 implies that $\pi^{(J)}$ does not fall when $\lambda_1^{(J)}, \ldots, \lambda_K^{(J)}$ undergo a MPS, which, by Lemma B5, happens when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. □

B3 Claims in Appendix Sections A2 and A3

B3.1 Proof of Proposition A1

Let $\mathcal{S}'$ be a superset of $\mathcal{S}$. Then

$$\begin{aligned} \mathbb{V}(\theta_k) &= \mathbb{E}[\mathbb{V}(\theta_k \mid \mathcal{S})] + \mathbb{V}(\mathbb{E}[\theta_k \mid \mathcal{S}]) \\ &\geq \mathbb{V}(\theta_k \mid \mathcal{S}) \\ &= \mathbb{E}[\mathbb{V}(\theta_k \mid \mathcal{S}, \mathcal{S}') \mid \mathcal{S}] + \mathbb{V}(\mathbb{E}[\theta_k \mid \mathcal{S}, \mathcal{S}'] \mid \mathcal{S}) \\ &\geq \mathbb{V}(\theta_k \mid \mathcal{S}, \mathcal{S}') \\ &= \mathbb{V}(\theta_k \mid \mathcal{S}') \end{aligned}$$

for each $k \in \{1, \ldots, K\}$, where the first two equalities hold by the law of total variance, the inequalities hold because the posterior variance of θ_k is non-negative and non-random (by Lemma 2), and the last equality holds because $\mathcal{S}'$ is a superset of $\mathcal{S}$. It follows that $0 \leq \pi(\mathcal{S}) \leq \pi(\mathcal{S}')$, thereby establishing (i) and (ii).

Now consider (iii). Differentiating the posterior variance matrix (15) with respect to σ_u^2 gives

$$\begin{aligned} \frac{\partial}{\partial \sigma_u^2} \mathbb{V}(\theta \mid \mathcal{S}) &= -\left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right) \left(\frac{\partial}{\partial \sigma_u^2} \left[\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right]\right) \left(\Sigma^{-1} + \frac{1}{\sigma_u^2} G\right) \\ &= \frac{1}{\sigma_u^4} \left(\Sigma^{-1} G \Sigma^{-1} + \frac{1}{\sigma_u^2} G^2 \Sigma^{-1} + \frac{1}{\sigma_u^2} \Sigma^{-1} G^2 + \frac{1}{\sigma_u^4} G^3\right), \end{aligned}$$

which is the sum of four matrices with strictly positive traces. Thus

$$\operatorname{tr}\left(\frac{\partial}{\partial\sigma_u^2}\,\mathbb{V}(\theta \mid \mathcal{S})\right) > 0$$

and therefore

$$\begin{aligned}\frac{\partial\pi(\mathcal{S})}{\partial\sigma_u^2} &= -\frac{1}{K}\operatorname{tr}\left(\frac{\partial}{\partial\sigma_u^2}\,\mathbb{V}(\theta \mid \mathcal{S})\right)\\ &< 0\end{aligned}$$

because traces are linear operators. □

B3.2 Proof of Proposition A2

Our proof of Proposition A2 uses the following fact about rank-one updates of invertible matrices.

Lemma B6 (Sherman-Morrison formula)**.** *Let $n \geq 1$ be an integer, let $A \in \mathbb{R}^{n\times n}$ be invertible, and let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$. If $v^T A^{-1} u \neq -1$, then*

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}.$$

Proof. See Bartlett (1951).

Proof of Proposition A2. Suppose $\mathcal{S}$ contains a single observation and let $w \equiv w^{(1)}$ for convenience. Then, by Lemmas 2 and B6, we have

$$\begin{aligned}\mathbb{V}(\theta \mid \mathcal{S}) &= \left(\Sigma^{-1} + \frac{1}{\sigma_u^2}ww^T\right)^{-1}\\ &= \Sigma - \frac{\Sigma ww^T\Sigma}{w^T\Sigma w + \sigma_u^2}.\end{aligned}$$

Thus

$$\begin{aligned}K\pi(\mathcal{S}) &= \operatorname{tr}(\Sigma - \mathbb{V}(\theta \mid \mathcal{S}))\\ &= \operatorname{tr}\left(\frac{\Sigma ww^T\Sigma}{w^T\Sigma w + \sigma_u^2}\right)\\ &= \frac{w^T\Sigma^2 w}{w^T\Sigma w + \sigma_u^2},\end{aligned}$$

where the last equality holds by the linearity and cyclic property of matrix traces. Equation (A2) follows. The inequalities (A3) follow from Proposition 1, as do the choices of $w^{(1)}$ that make $\star$ and $\star\star$ hold with equality. □

B3.3 Proof of Proposition A3

Define $g(z) \equiv nz^2/K(nz + K\sigma_u^2)$ for all $z > 0$. Then $g : (0, \infty) \to \mathbb{R}$ is convex. So if $\mathcal{S}$ is representative, then, by Lemma 3, its value

$$\pi(\mathcal{S}) = \sum_{k=1}^{K} g(\lambda_k)$$

does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. □

B3.4 Derivation of (A5)–(A7)

Derivation of (A5) *and* (A6). Let $k > R$. Now $\theta_1, \ldots, \theta_R, \theta_k$ are jointly normally distributed under the agent's prior, and so Lemma B1 implies that θ_k is conditionally normally distributed with mean

$$\mathbb{E}[\theta_k \mid \theta_{\mathcal{S}}] = \mathbb{E}[\theta_k] + \xi_k^T(\theta_{\mathcal{S}} - \mathbb{E}[\theta_{\mathcal{S}}]) \tag{B10}$$

and variance (A6) given $\theta_{\mathcal{S}}$. So

$$\begin{aligned}
\mathbb{V}(\theta_k \mid \mathcal{S}) &= \mathbb{V}(\mathbb{E}[\theta_k \mid \mathcal{S}, \theta_{\mathcal{S}}] \mid \mathcal{S}) + \mathbb{E}[\mathbb{V}(\theta_k \mid \mathcal{S}, \theta_{\mathcal{S}}) \mid \mathcal{S}] \\
&= \mathbb{V}(\mathbb{E}[\theta_k \mid \theta_{\mathcal{S}}] \mid \mathcal{S}) + \mathbb{E}[\mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) \mid \mathcal{S}] \\
&= \xi_k^T \, \mathbb{V}(\theta_{\mathcal{S}} \mid \mathcal{S}) \xi_k + \mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}),
\end{aligned}$$

where the first equality holds by the law of total variance, the second holds because θ_k is conditionally independent of $\mathcal{S}$ given $\theta_{\mathcal{S}}$, and the third uses (B10) and the non-randomness of (A6). So

$$\begin{aligned}
\operatorname{tr}(\mathbb{V}(\theta \mid \mathcal{S})) &= \sum_{k=1}^{K} \mathbb{V}(\theta_k \mid \mathcal{S}) \\
&= \sum_{k=1}^{R} \mathbb{V}(\theta_k \mid \mathcal{S}) + \sum_{k>R} \Big(\xi_k^T \, \mathbb{V}(\theta_{\mathcal{S}} \mid \mathcal{S}) \xi_k + \mathbb{V}(\theta_k \mid \theta_{\mathcal{S}}) \Big).
\end{aligned}$$

□

Derivation of (A7). Suppose θ has prior variance (8). Then

$$\mathbb{V}(\theta_{\mathcal{S}}) = (\rho \mathbf{1}_R \mathbf{1}_R + (1 - \rho) I_R)\sigma^2$$

has inverse

$$\begin{aligned}
\mathbb{V}(\theta_{\mathcal{S}})^{-1} &= \frac{1}{(1-\rho)\sigma^2} \left(I_R + \frac{\rho}{1-\rho} \mathbf{1}_R \mathbf{1}_R^T \right)^{-1} \\
&= \frac{1}{(1-\rho)\sigma^2} \left(I_R - \frac{\rho}{1 - \rho + \rho \mathbf{1}_R^T \mathbf{1}_R} \mathbf{1}_R \mathbf{1}_R^T \right)
\end{aligned}$$

by Lemma B6. Now if $k > R$, then

$$\begin{bmatrix} \Sigma_{11} \\ \vdots \\ \Sigma_{Rk} \end{bmatrix} = \rho \sigma^2 \mathbf{1}_R$$

and so

$$\begin{aligned}\xi_k^T \mathbb{V}(\theta_{\mathcal{S}})\xi_k &= \begin{bmatrix}\Sigma_{11}\\ \vdots \\ \Sigma_{Rk}\end{bmatrix}^T \mathbb{V}(\theta_{\mathcal{S}})^{-1} \begin{bmatrix}\Sigma_{11}\\ \vdots \\ \Sigma_{Rk}\end{bmatrix} \\ &= (\rho\sigma^2 \mathbf{1}_R)^T \left(\frac{1}{(1-\rho)\sigma^2} \left(I_R - \frac{\rho}{1-\rho+\rho \mathbf{1}_R^T \mathbf{1}_R} \mathbf{1}_R \mathbf{1}_R^T \right) \right) (\rho\sigma^2 \mathbf{1}_R) \\ &= \frac{\rho^2 R \sigma^2}{1+\rho(R-1)}\end{aligned}$$

because $\mathbf{1}_R^T \mathbf{1}_R = R$. Substituting this expression and $\mathbb{V}(\theta_k) = \sigma^2$ into (A6) yields (A7). □

B3.5 Proof of Proposition A4

It suffices to prove (ii) and (iii), which together imply (i).

If $\lambda_1, \ldots, \lambda_K$ are equal, then $\lambda_k = \overline{\lambda}$ for each $k \in \{1, \ldots, K\}$ and so

$$\begin{aligned}\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) &= -\frac{1}{2} \sum_{k=1}^{K} \ln(1) \\ &= 0,\end{aligned}$$

thus establishing (ii). For (iii), consider the function $g : (0, \infty) \to \mathbb{R}$ defined by

$$g(z) \equiv \frac{\ln(\overline{\lambda}) - \ln(z)}{2}.$$

This function is convex on its domain. So, by Lemma 3, the KL divergence

$$\mathcal{D}_{\mathrm{KL}}(\mathbb{P} \parallel \mathbb{P}^{(0)}) = \sum_{k=1}^{K} g(\lambda_k)$$

from $\mathbb{P}$ to $\mathbb{P}^{(0)}$ does not fall when $\lambda_1, \ldots, \lambda_K$ undergo a MPS. □