Robust Latent Data Representations*

Larry Samuelson
Yale University

Jakub Steiner
University of Zurich, CERGE-EI, and CTS

June 23, 2026

Abstract

We study how an empirical Bayesian applies a latent-type theory to an observed sample. She starts with primitive types, each a distribution over observables, but may find the sample inconsistent with any mixture of the types. We axiomatically characterize a familiar statistical procedure—maximum-likelihood estimation followed by Bayesian updating—that adjusts the type definitions to match the sample. This procedure endogenizes latent types, usually primitive in economic theory, and addresses the wedge between predictions and observations left open by Berk-Nash reasoning. Economic examples illustrate the interplay between data and the agent's view of the latent types.

*This paper builds on and supersedes our previous paper, "Constrained Data-Fitters." We thank Sandro Ambuehl, Stéphane Bonhomme, Mira Frick, Heidi Thysen, Ryota Iijima, Emir Kamenica, Giacomo Lanzani, Rava da Silveira, Ran Spiegler, Colin Stewart, Michael Woodford, Andriy Zapechelnyuk, and various seminar and workshop audiences for comments. We thank Pavel Kocourek for research assistance. Steiner has benefited from grant GAČR 24-10145S.

1 Introduction

Economic agents often use statistical theories to reason about the latent characteristics underlying the data they observe. An employer who observes her workers' outputs may use a statistical theory to draw inferences about their hidden efforts. An investor who observes the investment positions of her peers may use a theory to reason about their hidden market sentiments. These analysts not only form a *generative model* of the process that generated the data, but also form a joint belief about the observed and latent characteristics of the realized observations themselves.

The distinction is straightforward but important. A generative model is useful for reasoning about a new worker or new investor drawn from the general population. But an employer who reasons about her existing workforce, or an investor who reasons about her peers, needs a belief that both matches the sample's empirical distribution of observables and assigns latent characteristics to the observations in that sample. We call such a belief a *sample model*. Together, the generative model and the sample model constitute a *latent representation*.

The distinction between a generative model and a sample model becomes substantive when the agent's statistical theory is misspecified, because the observed sample may then be inconsistent with the agent's theory. For example, an employer using a misspecified theory of the labor force may find that no distribution over hidden efforts can reproduce the output distribution observed in her workforce. Unlike her reasoning about new workers drawn from a general population, her reasoning about her current workforce must confront the wedge between observations and theory.

We study how a misspecified analyst constructs a latent representation of the data she observes. She begins with a theory consisting of primitive latent types, each describing a distribution over observable outcomes. As in statistical mixture models, she estimates the prevalence of these types from the observed sample. For instance, an employer's theory may specify the output distributions generated by low and high effort, and she may use the observed output distribution to estimate the prevalence of each effort level in her workforce. Under misspecification, however, no mixture of the primitive types may match the sample, and we ask how the analyst adjusts the type definitions to account for the mismatch. We therefore study a form of statistical reasoning in which the latent types themselves become endogenous. This separates our approach from Berk-Nash reasoning, which identifies a best-fitting generative model but leaves open how the analyst reasons about the realized sample itself.

A natural way to construct a latent data representation is a two-stage empirical Bayes procedure. In the first stage, as in a latent-type mixture model, the analyst estimates the distribution over primitive types by maximum likelihood. This step produces a fitted generative model: the analyst's best account, within her primitive theory, of the process that generated the observed data. In the

second stage, the analyst applies Bayes' rule within this fitted generative model to assign posterior probabilities over latent types to each observed outcome, and then combines these posterior assignments with the empirical distribution of outcomes in the realized sample to obtain a sample model. The construction is close in spirit to misspecified learning in Berk-Nash equilibrium, but addresses a different question. Berk-Nash reasoning identifies the best-fitting generative model; our sample model describes how the analyst reasons about the realized sample when that fitted model does not reproduce its empirical distribution.

Our main result provides an axiomatic characterization of this MLE-Bayes procedure. The axioms do not ask the analyst to maximize a likelihood or to apply Bayes' rule, nor do they refer to a generative model. Nonetheless, Theorem 1 states that these restrictions characterize the MLE-Bayes procedure. Any sample model satisfying the axioms can be represented as if the analyst had formed it by this procedure; conversely, the MLE-Bayes sample model satisfies the axioms.

The axioms make reference only to the sample model. A "realism" axiom requires the sample model's observable marginal to match the empirical distribution in the sample, and two "conservatism" axioms require, to the extent possible, the sample model to stay true to the analyst's theory. If the data can be explained as a mixture of the latent types in the analyst's theory, the sample model does so. When the data cannot be so explained, the sample model still preserves the primitive likelihood ratios across employed types, and hence the diagnostic content of each observation. The familiar MLE-Bayes procedure is therefore the unique way to combine realism about the sample with conservatism toward the primitive definitions of latent types.

Theorem 1 clarifies what the analyst changes, and what she leaves unchanged, when the data cannot be matched by her theory. She changes the predictive content of the types: the sample types are adjusted so that their mixture reproduces the observed sample. But she preserves diagnostic content: if a labor outcome was relatively more indicative of high than low effort under the employer's primitive theory, it remains so in the sample model. Theorem 2 sharpens this interpretation by extending this conservatism across samples. Once the analyst has formed sample types, she preserves their definitions after a sample shift as long as the new sample can still be represented as a mixture of those types. Only when realism makes this impossible does she redefine the types. Thus, the procedure is conservative both relative to the primitive theory and across samples.

We illustrate the results in three applications. The first asks which elements of a misspecified theory—interpreted as stereotypes in a labor-market context—are corrected by the empirical evidence contained in the sample, and which remain rigid under the MLE-Bayes procedure. Realism forces a correction of the predictive content of the stereotypes: when the sample shows that a given demographic group is more productive than the primitive theory asserted, the MLE-Bayes analyst incorporates this fact into her sample model. The procedure, however, leaves the diagnostic content

of the stereotypes intact. If the primitive theory attributed high labor output among members of a certain demographic group to luck rather than talent, then the empirical evidence does not by itself reverse this interpretation.

The second application explores a principal-agent relationship, contrasting three versions of the principal. We begin with a benchmark principal who correctly understands the output distributions generated by the effort levels. We then examine a misspecified principal who reasons in the spirit of Berk-Nash equilibrium, using the fitted generative model to guide her contract choice. Because this principal evaluates incentives through the primitive effort-output distributions, her contract reflects the misspecified technology and differs from the benchmark contract. Finally, we examine a misspecified principal whose contract choice is guided by the sample model, and who therefore lets the observed sample reshape her perceived production technology. This principal duplicates the benchmark contract, but not the benchmark outcome. Equilibrium feedback between effort, observed output, and the principal's interpretation of the technology prevents her from implementing full effort. Thus, two principals may write the same contract, yet induce different outcomes because they interpret effort-output data differently.

The third application illustrates Theorem 2. We let the analyst observe a sample of choices in a population and infer the underlying distribution of hidden characteristics, as when an investor infers her peers' market sentiments from their positions. The theorem implies that the analyst preserves her sample-type definitions whenever the new sample can still be represented as a mixture of the current sample types. Thus, when a shift in aggregate behavior can be explained without type redefinitions, the inferred distribution of types reacts strongly to the aggregate shift: because the meanings of the types are kept fixed, the change in observed behavior is attributed fully to a changing distribution of types in the population.

The economics of incomplete information has long taken types to be exogenously and correctly specified. The recent literature on misspecified learning relaxes the second assumption. We complement this literature by relaxing the first assumption. Rather than focusing only on how the agent reasons about the process that generated the data, we study how she represents the realized data themselves, and in doing so we endogenize the latent types. We close the paper by relating this perspective to work in machine learning, variational Bayesian statistics, and rational inattention.

2 Samples, Types, and Sample Models

An analyst observes a sequence of outcomes x^i, $i = 1, \dots, n$, each taking a value in a finite set $\mathcal{X}$. For the purposes of our analysis, the sequence is fully described by the empirical frequencies

$$s(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x^i = x}, \quad x \in \mathcal{X}.$$

We refer to the sequence of outcomes and its empirical distribution s interchangeably as the *sample*. We allow for any sample size, and let s be any distribution in the interior of $\Delta(\mathcal{X})$, abstracting from the restriction imposed by finite n.

The analyst is endowed with a theory of the process that generated the sample. According to this theory, each outcome x^i was drawn from an outcome distribution $\pi_{z^i} \in \operatorname{int} \Delta(\mathcal{X})$, where z^i is a latent type label from a finite set $\mathcal{Z}$.[1] We refer to the distributions π_z as *primitive types*, and to the collection of primitive types $\mathcal{T} = \{\pi_z\}_{z \in \mathcal{Z}}$ as a *theory*. The theory specifies the conditional distributions π_z, but not the prevalence of the latent values; the analyst treats the distribution over $\mathcal{Z}$ as an unrestricted parameter to be inferred from the sample. We assume that the primitive types $\pi_z \in \mathcal{T}$ are affinely independent.[2] See Figure 1A.

The analyst forms a belief about a representative element (x^i, z^i) of the sample, obtained by drawing the index i uniformly from $1, \dots, n$. We denote this belief by $q^s_{XZ} \in \Delta(\mathcal{X} \times \mathcal{Z})$, and refer to it as the analyst's *sample model*; $q^s_{XZ}(x, z)$ is the probability that the analyst assigns to the event that the index i, drawn uniformly from $1, \dots, n$, satisfies $(x^i, z^i) = (x, z)$. We denote the associated marginal and conditional distributions by q^s_X, q^s_Z, $q^s_{X|Z=z}$, and $q^s_{Z|X=x}$.

Example 1. The analyst is a principal who observes the outputs $x^i \in \mathcal{X}$ of her workers $i = 1, \dots, n$. Her theory $\mathcal{T} = \{\pi_\ell, \pi_h\}$ states that each worker i has exerted either low or high effort and that an effort $z \in \mathcal{Z} = \{\ell, h\}$ generates an outcome by drawing from $\pi_z \in \operatorname{int} \Delta(\mathcal{X})$. After observing the sample s of her workers' outcomes, the analyst forms a sample model q^s_{XZ} that assigns probabilities $q^s_{XZ}(x, z)$ to the event that a worker i, drawn uniformly from her workforce, has exerted effort $z^i = z$ and generated output $x^i = x$. In Section 5.2, we let the principal choose workers' incentives based on her sample model, let the observed sample be generated by workers who optimally respond to the incentives, and examine the fixed point. △

[1] The full-support assumption on the primitive types π_z is not essential, but avoids dealing with zero-probability events.

[2] Affine independence holds when a statistician works with an identified mixture model. The assumption implies uniqueness of the analyst's sample model studied below.

3 Model Formation

We present four axioms governing how the analyst combines her sample s with her theory $\mathcal{T}$ to form a sample model q^s_{XZ}. Our main result shows that these axioms characterize a familiar two-stage MLE-Bayes statistical procedure. The procedure first estimates the type mixture by maximum likelihood and then uses Bayes' rule to assign latent types to observed outcomes.

3.1 Axioms

The analyst forms a model of her sample, rather than of the data-generating process. Accordingly, her sample model must match the observed sample exactly:

A1 (realism): $q^s_X = s$.

For instance, an employer's model of her representative worker must match the observed distribution s of her workers' outcomes.

We call q^s_Z the *sample prior*. For each employed type $z \in \operatorname{supp} q^s_Z$, we call $q^s_{X|Z=z}$ the *sample type*. We distinguish these sample types from the primitive types π_z, and view the sample types as the analyst's reassessment of her primitive types after observing the sample. When $s \notin \operatorname{co}(\mathcal{T})$, such reassessment is forced by realism (A1).

For a finite sample size n, the empirical distribution s may lie outside the convex hull of the primitive types because of sampling fluctuations. More importantly, if the primitive types π_z are misspecified, the limiting empirical distribution may itself lie outside $\operatorname{co}(\mathcal{T})$. We focus on this large-sample interpretation, in which the failure of exact fit reflects misspecification rather than sampling noise. For instance, the principal from Example 1, if endowed with a misspecified theory of how effort levels generate outputs, may observe an output distribution $s \notin \operatorname{co}(\mathcal{T})$ with probability approaching one as the size of her workforce diverges. Then, to reason about a representative worker's latent effort level while satisfying realism, she must replace the primitive effort-output distributions π_z with sample types $q^s_{X|Z=z}$.

A general adjustment of the type definitions could take the form

$$q^s_{X|Z=z}(x) = \pi_z(x) d_z(x)$$

where the "discrepancy function" $d_z(x)$ allows the wedge between $q^s_{X|Z=z}(x)$ and $\pi_z(x)$ to depend on both the latent type z and the outcome x. However, the analyst observes only the marginal distribution of X, not the latent types, and so the sample does not by itself inform how the discrepancy should be allocated across latent types. Axiom A2 imposes a conservative resolution: the analyst

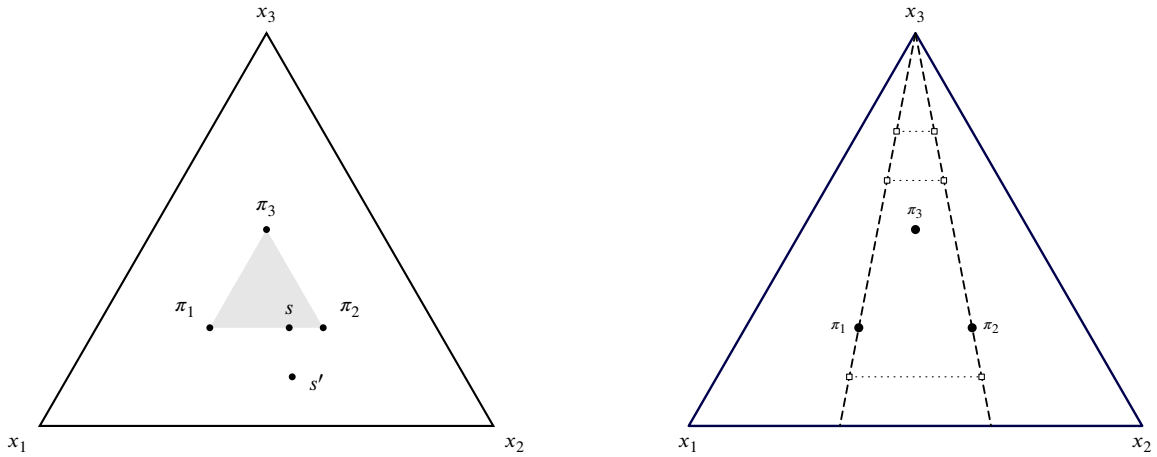


Figure 1: Display A: Primitive latent types π_z, $z \in \mathcal{Z} = \{1, 2, 3\}$ and two samples s and s'. Display B: For a fixed set of employed types, here assumed to be $\operatorname{supp} q_Z^s = \{1, 2\}$, A2 imposes a linear condition on the pair $(q_{X|Z=1}^s, q_{X|Z=2}^s)$: the two sample types must lie on the two dashed lines, paired as illustrated by the three example pairs depicted by white squares.

uses a common outcome-specific adjustment, $d_z(x) = d(x)$, for all employed types. This adjustment changes the predictive content of the types while preserving their likelihood ratios, and hence their diagnostic meanings:

A2 (diagnostic conservatism):

$$\frac{q_{X|Z=z}^s(x)}{q_{X|Z=z'}^s(x)} = \frac{\pi_z(x)}{\pi_{z'}(x)} \quad \text{for all } z, z' \in \operatorname{supp} q_Z^s \text{ and all } x \in \mathcal{X}. \tag{1}$$

Axiom A2 states that the analyst preserves her view of the relative evidence that outcome x provides in favor of one type over another. For instance, in our running example, the sample may force the principal to revise her views on the output distribution generated by each effort z, but she maintains her view of how indicative each output is of high rather than low effort.

Fixing the set of types in the support of the sample model, A2 imposes a linear restriction on the tuple of sample types $(q_{X|Z=z}^s)_{z \in \operatorname{supp} q_Z^s}$. Figure 1B illustrates.

The remaining two axioms pertain to the analyst's choice of the sample prior q_Z^s. As in the empirical Bayes approach pioneered by Robbins (1956), the analyst estimates the prior distribution q_Z^s over types by calibrating her theory to match the observed sample, allowing any prior in $\Delta(\mathcal{Z})$.

Axiom A3 states that the analyst is dogmatic about her theory whenever realism (A1) permits: she maintains the primitive types π_z whenever her theory allows a perfect fit, as for the sample s in Figure 1A.

A3 (perfect-fit conservatism): If $s \in \text{co}(\mathcal{T})$, then $q^s_{X|Z=z} = \pi_z$ for all $z \in \text{supp}\, q^s_Z$.

Thus, using the affine independence of $\mathcal{T}$, whenever $s \in \text{co}(\mathcal{T})$, Axioms A1 and A3 uniquely pin down the sample prior q^s_Z, and hence the sample model $q^s_{XZ}(x,z) = q^s_Z(z)\pi_z(x)$.

The final axiom requires the analyst to extend these perfect-fit models continuously.

A4 (continuity): The map $s \mapsto q^s_Z$ is continuous on $\text{int}\,\Delta(\mathcal{X})$.

A4, together with A1 and A2, extends the perfect-fit identification of the sample model to samples that the primitive theory cannot match.

3.2 MLE-Bayes Procedure

In preparation for a representation theorem, we describe a familiar two-stage procedure for constructing sample models, based on maximum-likelihood estimation and Bayes' rule. In the first stage, the analyst forms the maximum-likelihood estimate of the data-generating distribution. In the second stage, she combines this estimate with the observed sample to construct the sample model.

In the first stage, the analyst assumes that the observed sample is generated by i.i.d. draws from a latent-type mixture model with primitive conditionals $p_{X|Z=z} = \pi_z$, and estimates the prior p_Z. Let p^s_Z be the maximum-likelihood estimate formed upon observing the sample s—that is, the unique maximizer in[3]

$$\max_{p_Z \in \Delta(\mathcal{Z})} \sum_{x \in \mathcal{X}} s(x) \ln \left(\sum_{z \in \mathcal{Z}} p_Z(z)\pi_z(x) \right). \tag{2}$$

We let

$$p^s_{XZ}(x,z) := \pi_z(x)p^s_Z(z)$$

denote the implied joint distribution and refer to it as the *generative model*.

In the second stage, the analyst constructs the sample model by applying Bayes' rule to the generative model. For an observation i uniformly drawn from $1, \ldots, n$, she assigns the conditional probability that the latent type Z^i equals z given $X^i = x$ according to

$$q^{\mathsf{MLE},s}_{Z|X=x}(z) := p^s_{Z|X=x}(z) \quad \text{for all } (x,z) \in \mathcal{X} \times \mathcal{Z}.$$

She then combines this conditional belief with the observed sample to form the *MLE-Bayes sample model*

$$q^{\mathsf{MLE},s}_{XZ}(x,z) = s(x)p^s_{Z|X=x}(z) \quad \text{for all } (x,z) \in \mathcal{X} \times \mathcal{Z}. \tag{3}$$

[3] Affine independence implies that the objective in (2) is strictly concave, and thus its maximizer is unique.

Using this MLE-Bayes procedure, the analyst forms two joint distributions, p^s_{XZ} and $q^{\mathsf{MLE},s}_{XZ}$. We call the pair $(p^s_{XZ}, q^{\mathsf{MLE},s}_{XZ})$ a *latent representation*. A wedge between these two joint distributions arises whenever the analyst's generative prediction p^s_X differs from the observed sample s.

The generative model p^s_{XZ} and the sample model $q^{\mathsf{MLE},s}_{XZ}$ represent the analyst's out-of-sample and within-sample beliefs, respectively. Returning to our running example and assuming that the principal's workers are drawn from a general population, $p^s_{XZ}(x, z)$ is the probability the analyst assigns to the event that a worker newly drawn from the general population has type z (i.e., exerts effort level z) and generates outcome x, whereas $q^{\mathsf{MLE},s}_{XZ}(x, z)$ is the principal's belief about the representative worker from her current workforce.

4 Results

4.1 Partial Characterization

Before stating our main characterization theorem, we begin with a partial characterization of the sample model based only on Axioms A1 and A2. These two axioms do not determine which types the analyst employs. They do, however, determine the sample model once the employed support is fixed.

For a nonempty set $\mathcal{E} \subseteq \mathcal{Z}$, let $q^{\mathsf{MLE},s,\mathcal{E}}_{XZ}$ be the MLE-Bayes sample model formed in the MLE-Bayes procedure but with the maximum-likelihood estimation (2) restricted to priors p_Z in $\Delta(\mathcal{E})$, where types outside $\mathcal{E}$ receive probability zero. The next proposition shows that A1 and A2 leave open which support is employed, but impose the restricted MLE-Bayes representation once that support is fixed.

Proposition 1 (Fixed-support representation)**.** *Fix an interior sample s and a nonempty set $\mathcal{E} \subseteq \mathcal{Z}$. There exists at most one sample model q^s_{XZ} satisfying A1 and A2 with*

$$\operatorname{supp} q^s_Z = \mathcal{E}.$$

If such a sample model exists, then $q^s_{XZ} = q^{\mathsf{MLE},s,\mathcal{E}}_{XZ}$.

Once the employed support is known, A1 and A2 pin down the analyst's sample model as the restricted MLE-Bayes model. The remaining task is to select the support. Axiom A3 does so at perfect-fit samples, where the primitive theory itself can match the data. Axiom A4 then extends this support-selection rule away from perfect-fit samples.

4.2 Full Characterization

Our main result establishes that A1–A4 characterize the MLE-Bayes procedure. The axioms make no reference to a generative model; they speak only to how the analyst forms a model of the observed sample. Nonetheless, any analyst satisfying them can be represented as if she first estimated a generative latent-type mixture model by maximum likelihood and then used Bayes' rule to convert that estimate into a model of the observed sample. Conversely, the MLE-Bayes procedure satisfies A1–A4. In particular, it satisfies diagnostic conservatism: even when the observed sample cannot be matched by the primitive theory, and the analyst therefore adjusts the predictive content of the types, the procedure preserves the primitive likelihood ratios across employed types.

Theorem 1. *The map $s \mapsto q^s_{XZ}$ satisfies A1–A4 if and only if $q^s_{XZ} = q^{\mathsf{MLE},s}_{XZ}$ for every sample s.*

Since the maximum-likelihood estimate exists, and it is unique when the primitive latent types are affinely independent, the result implies that Axioms A1–A4 identify a unique sample model q^s_{XZ} for each interior sample s. From now on, we identify q^s_{XZ} with $q^{\mathsf{MLE},s}_{XZ}$.

4.3 Type Conservatism

Our second result shows that, after a sample shift, the analyst preserves her sample-type definitions whenever realism (A1) permits. Suppose that the sample s gives rise to the sample model q^s_{XZ}, and consider a new sample s'. If the analyst is to retain the current sample types $q^s_{X|Z=z}$, then realism requires s' to lie in their convex hull. The theorem below states that this necessary condition is also sufficient. Hence the analyst retains the current type definitions for every sample that can be represented as a mixture of them. She redefines the sample types only when such a representation is impossible.

We say that the analyst *retains her type definitions* in a sample shift from s to s' if

$$\operatorname{supp} q^{s'}_Z \subseteq \operatorname{supp} q^s_Z, \text{ and } q^{s'}_{X|Z=z} = q^s_{X|Z=z} \text{ for all } z \in \operatorname{supp} q^{s'}_Z.$$

Theorem 2 (Maximal type conservatism)**.** *The following statements are equivalent:*

1. *The analyst retains her type definitions in the sample shift from s to s'.*
2. $s' \in \operatorname{co}\left(\{q^s_{X|Z=z}\}_{z \in \operatorname{supp} q^s_Z}\right)$.

4.4 Proof Outlines

We start by characterizing the maximum-likelihood estimate from the first stage of the MLE-Bayes procedure. For this, we introduce a *discrepancy vector*,

$$d(x) = \frac{s(x)}{p_X^s(x)} \quad \text{for all } x \in \mathcal{X},$$

which expresses the wedge between the observed sample s and the analyst's generative prediction p_X^s. The following lemma, which follows from the first-order conditions for the maximum-likelihood estimation, states that the generative model employs those types that provide the best explanation of this discrepancy.

For $a, b \in \mathbb{R}^{\mathcal{X}}$ define the scalar product $a \cdot b := \sum_{x \in \mathcal{X}} a(x)b(x)$.

Lemma 1 (MLE optimality conditions)**.** *The maximum-likelihood estimate p_Z^s satisfies the necessary and sufficient conditions*

$$\pi_z \cdot d = 1 \qquad \textit{for all } z \in \operatorname{supp} p_Z^s, \tag{4}$$

$$\pi_z \cdot d \leq 1 \qquad \textit{for all } z \notin \operatorname{supp} p_Z^s. \tag{5}$$

The next result, which follows from these optimality conditions, states that latent representations satisfy two coherence properties.

Lemma 2 (Coherence of latent representations)**.** *Let $(p_{XZ}^s, q_{XZ}^{\mathsf{MLE},s})$ be the latent representation for sample s. Then the following hold:*

(i) Rational expectations: $q_Z^{\mathsf{MLE},s} = p_Z^s$.

(ii) Diagnostic conservatism: $q_{XZ}^{\mathsf{MLE},s}$ *satisfies A2.*

The rational-expectations property states that when the analyst compares the estimated distribution $q_Z^{\mathsf{MLE},s}$ of the latent types in the observed sample with the estimated generative prior p_Z^s, then no surprise arises: $q_Z^{\mathsf{MLE},s} = p_Z^s$.[4]

We then use Lemmas 1 and 2 to establish Proposition 1: any sample model with a given support that satisfies A1–A2 must coincide with the MLE-Bayes sample model restricted to that support. The proof of Theorem 1 adds A3 and A4 to identify the support.

[4]The result formalizes an intuition sometimes invoked in support of rational expectations: when expectations and outcomes diverge, the wedge should be removed by adjusting expectations. Indeed, in our framework, should a wedge between p_Z^s and $q_Z^{\mathsf{MLE},s}$ arise, aligning p_Z^s with $q_Z^{\mathsf{MLE},s}$ increases the generative model's likelihood.

Axiom A3 identifies the sample model at perfect-fit samples. To extend the identification to a sample $s \notin \text{co}(\mathcal{T})$, the proof connects s to the fitted observable distribution p^s_X by the path

$$s_t = ts + (1-t)p^s_X, \qquad t \in [0,1].$$

At $t = 0$, the sample $s_0 = p^s_X$ is a perfect-fit sample, so the axiomatic sample model coincides with the MLE-Bayes sample model. Moreover, we show that the MLE prior is constant along this path. A continuation argument, using A4 and Proposition 1, then propagates this coincidence from $t = 0$ all the way to $t = 1$. Since $s_1 = s$, the axiomatic sample model at s is the MLE-Bayes sample model. The converse follows by directly verifying A1–A3 and using continuity of the unique maximum-likelihood estimate to obtain A4.

The proof of Theorem 2, also in Appendix A, uses the MLE optimality conditions and the coherence properties of latent representations. If the new sample s' lies in the convex hull of the current sample types, then s' can be obtained by changing only the sample prior while keeping the conditional distributions $q^s_{X|Z=z}$ fixed. The proof shows that this candidate sample model is precisely the MLE-Bayes sample model at s'.

5 Applications

5.1 Stereotypes

We begin with a simple illustration of Theorem 1 that highlights a limitation of the MLE-Bayes procedure as a mechanism for correcting stereotypes. One might hope that sufficiently strong empirical evidence would force the analyst to abandon stereotypical beliefs. The example shows that this hope is only partly warranted. The procedure corrects the predictive content of a stereotype, in accordance with realism (A1), but preserves the stereotype's diagnostic content, in accordance with A2.

The analyst observes $x^i = (x^i_1, x^i_2)$ for each individual i, where $x^i_1 \in \{L, H\}$ is labor output and $x^i_2 \in \{0, 1\}$ is a demographic group. The analyst's primitive theory $\mathcal{T} = \{\pi_\ell, \pi_h\}$ consists of outcome distributions π_ℓ and π_h for low- and high-talent types:

	$(L,0)$	$(H,0)$	$(L,1)$	$(H,1)$
π_ℓ	$\frac{3}{5}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{10}$
π_h	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{2}$

.

Under this theory, high-talent types are more likely to produce high output than low-talent types, and group 0 is more prevalent among the low types and group 1 is more prevalent among the high

types. For $x = (x_1, x_2)$, the primitive likelihood ratio $\lambda(x) := \pi_h(x)/\pi_\ell(x)$ satisfies

$$\lambda(L, 0) = \lambda(H, 0) = \frac{1}{2}, \tag{6}$$

whereas

$$1 = \lambda(L, 1) < \lambda(H, 1) = 5. \tag{7}$$

Thus, the primitive theory indicates that labor output is not diagnostic of talent for group 0: high and low output have the same likelihood ratio, as if output differences within this group reflected luck rather than talent. By contrast, for group 1, the theory views high output as strong evidence of high talent.

The analyst then encounters the sample $s = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, indicating that labor output and demographic group are independent in the observed population. This sample cannot be generated by any mixture of π_ℓ and π_h because both primitive types assign too little probability to high-output group-0 observations. The sample is thus inconsistent with the analyst's stereotypical primitive theory.

Straightforward calculations in Appendix B.1 show that the analyst forms the sample types

	$(L,0)$	$(H,0)$	$(L,1)$	$(H,1)$
$q^s_{X\|Z=\ell}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{12}$
$q^s_{X\|Z=h}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{5}{12}$

.

The sample thus corrects the predictive content of the primitive theory: the analyst adjusts her view of both low- and high-talent types, assigning more probability to high-output group-0 individuals than the primitive theory did. However, this correction is partial: the MLE-Bayes procedure preserves the diagnostic content of observations. By Theorem 1, it satisfies diagnostic conservatism (A2), and thus the sample types inherit (6) and (7):

$$\frac{q^s_{X|Z=h}(H,0)}{q^s_{X|Z=\ell}(H,0)} = \frac{q^s_{X|Z=h}(L,0)}{q^s_{X|Z=\ell}(L,0)} = \frac{1}{2};$$

that is, for group 0, labor output remains non-diagnostic of talent. By contrast, labor output remains strongly diagnostic for group 1:

$$1 = \frac{q^s_{X|Z=h}(L,1)}{q^s_{X|Z=\ell}(L,1)} < \frac{q^s_{X|Z=h}(H,1)}{q^s_{X|Z=\ell}(H,1)} = 5.$$

Thus, the MLE-Bayes procedure absorbs the mismatch with the sample by adapting the types' predictive content while preserving their diagnostic content.

5.2 Moral Hazard

We now expand on the moral-hazard environment introduced in Example 1 in which the principal uses the sample generated by workers' behavior to interpret the effort-output technology. The application highlights the distinction between reasoning from the generative model and reasoning from the sample model. A misspecified principal guided by the *generative* model, in the spirit of Berk-Nash equilibrium, responds to the fixed primitive technology, and therefore chooses a suboptimal contract. A misspecified principal guided by the *sample* model instead lets the observed sample reshape how she perceives the technology. In the setting studied below, this latter principal chooses the same contract as the well-specified benchmark principal. Still, misspecification imposes a cost: in equilibrium, she implements only an interior level of effort.

A unit mass of workers exerts either low or high effort. Effort $z \in \mathcal{Z} = \{\ell, h\}$ generates a stochastic outcome $x \in \mathcal{X} = \{L, H, N\}$ according to the objective distribution $\rho_z \in \Delta(\mathcal{X})$. We refer to $\rho := \{\rho_z\}_{z\in\mathcal{Z}}$ as the *objective technology*. Worker i receives wage $w(x^i)$ upon the realization of outcome x^i. The cost of high effort is $c_h > 0$, while low effort is costless, $c_\ell = 0$. Given the wage profile, workers choose a possibly mixed effort

$$e^* \in \mathrm{BR}\,(w, \rho) := \arg\max_{e\in\Delta(\mathcal{Z})} \sum_{(x,z)\in\mathcal{X}\times\mathcal{Z}} (w(x) - c_z)e(z)\rho_z(x), \tag{8}$$

knowing the objective technology ρ.[5] Risk neutrality allows us to abstract from risk-sharing and focus instead on the principal's subjective perception of the technology.

The principal, constrained by limited liability, commits to a nonnegative wage profile $w \in \mathbb{R}^{\mathcal{X}}_+$. She derives revenue $r(x)$ from outcome $x \in \mathcal{X}$ and chooses w to maximize expected revenue net of the wage expenditure. We compare three principals: a well-specified benchmark principal who knows the objective technology, a misspecified principal guided by the generative model, and a misspecified principal guided by the sample model. The contrast between the two variants of the misspecified principal highlights the role of the sample model.

The benchmark principal knows the objective technology ρ. We assume that high effort generates higher expected revenue than low effort,

$$\mathrm{E}_{\rho_h}\, r(x) > \mathrm{E}_{\rho_\ell}\, r(x),$$

[5]The workers' knowledge of the objective technology is motivated by the assumption that they, unlike the principal, observe efforts and outputs jointly.

and that H is the outcome most diagnostic of high effort,

$$\frac{\rho_h(H)}{\rho_\ell(H)} > \max\left\{\frac{\rho_h(L)}{\rho_\ell(L)}, \frac{\rho_h(N)}{\rho_\ell(N)}\right\}.$$

By standard arguments, when c_h is sufficiently low, the benchmark principal optimally implements high effort by rewarding only H, leaving the worker indifferent. She chooses

$$w_{\text{b}}(H) = \frac{c_h}{\rho_h(H) - \rho_\ell(H)} > 0 \quad \text{and} \quad w_{\text{b}}(L) = w_{\text{b}}(N) = 0, \tag{9}$$

with $w_{\text{b}}(H)$ implied by the binding incentive-compatibility constraint.

The two misspecified principals share the same primitive theory of the technology, $\mathcal{T} = \{\pi_z\}_{z\in\mathcal{Z}}$. Each type $\pi_z \in \Delta(\mathcal{X})$ specifies the principal's subjective initial view of how effort z generates the outcome X. We assume that the theory $\mathcal{T}$ is not too distorted. Specifically,

$$\frac{\pi_h(H)}{\pi_\ell(H)} > 1 > \frac{\pi_h(L)}{\pi_\ell(L)} \quad \text{and} \quad \frac{\pi_h(N)}{\pi_\ell(N)} = 1.$$

Thus, $\mathcal{T}$ views H as diagnostic of high effort, L as diagnostic of low effort, and N as indicative of neither effort. Finally, both misspecified principals prefer H to L, $r(H) > r(L)$, and therefore, like the benchmark principal, prefer high effort to low effort when c_h is sufficiently small.

The first misspecified principal reasons in the spirit of Berk-Nash equilibrium. Based on the sample s of outcomes observed in past interactions with workers, she estimates a *generative* model $p^s_{XZ}(x,z) = p^s_Z(z)\pi_z(x)$, and lets this model guide her choice of the wage profile. Thus, when choosing the wage profile, the principal assumes that effort z generates output according to $p^s_{X|Z=z}$. Since $p^s_{X|Z=z} = \pi_z$, the sample affects only the prior over efforts, not the principal's perceived effort-output technology. The resulting contract is therefore independent of the sample and rigidly informed by the primitive theory $\mathcal{T}$. As with her benchmark counterpart, when c_h is sufficiently low, this principal subjectively implements high effort by rewarding H:

$$w_{\text{g}}(H) = \frac{c_h}{\pi_h(H) - \pi_\ell(H)} > 0, \qquad w_{\text{g}}(L) = w_{\text{g}}(N) = 0.$$

Generically, $w_{\text{g}}(H) \neq w_{\text{b}}(H)$, with the wedge reflecting the misspecification $\pi_z \neq \rho_z$.

Our focus is the second misspecified principal, who instead best responds to the sample model. Assuming that the principal-worker interaction with a fixed labor force is repeated, we let the prin-

cipal observe the aggregate workers' output distribution,

$$s = \sum_{z \in \mathcal{Z}} e(z)\rho_z, \tag{10}$$

which is the sample induced by the workers' mixed effort e and the objective technology ρ. This principal engages with the wedge between the prediction of her generative model and the observed sample of labor outputs. Given s, the principal forms a sample model q^s_{XZ} satisfying A1–A4. Specifically, she believes that effort $z \in \operatorname{supp} q^s_Z$ generates output according to the sample type $q^s_{X|Z=z}$ instead of π_z.

We assume that the objective technology is less extreme than the principal's primitive theory in its diagnostic odds:

$$\frac{\pi_\ell(H)}{\pi_\ell(L)} < \frac{\rho_z(H)}{\rho_z(L)} < \frac{\pi_h(H)}{\pi_h(L)}, \quad z \in \{\ell, h\}. \tag{11}$$

Lemma 3 in Appendix B.2 establishes that, under this condition, the principal's sample model employs both types.

We denote the principal's *subjective technology* by $\hat{\rho}^s := \{q^s_{X|Z=\ell}, q^s_{X|Z=h}\}$. Given $\hat{\rho}^s$, the principal chooses the wage profile

$$w^* \in \arg\max_{w \in \mathbb{R}^{\mathcal{X}}_+} \max_{e \in \mathrm{BR}(w, \hat{\rho}^s)} \sum_{(x,z) \in \mathcal{X} \times \mathcal{Z}} (r(x) - w(x)) e(z) q^s_{X|Z=z}(x). \tag{12}$$

Specifically, the principal mistakenly anticipates that workers react to the wage profile according to the subjective technology $\hat{\rho}^s$.

A *sample equilibrium* is a fixed point (e^*, s^*, w^*) satisfying:

1. given w^*, the workers' mixed effort e^* is optimal according to (8),
2. the sample s^* is generated by the workers' effort e^* according to (10), and
3. given the subjective technology $\hat{\rho}^{s^*}$, the principal optimizes the wage profile w^* as in (12).

We interpret the equilibrium as a steady state of a repeated interaction with a fixed workforce. After observing the output distribution generated by a contract, the principal treats the wedge between her model's prediction and the observed sample as persistent rather than transitory. She therefore uses the resulting current sample types as her subjective effort-output technology when optimizing the contract.

Abusing notation, let $\hat{\rho}^z = \{q^{\rho_z}_{X|Z=\tilde{z}}\}_{\tilde{z} \in \mathcal{Z}}$ denote the subjective technology formed by the principal when all workers exert effort z, so that she observes sample $s = \rho_z$. We impose a key mono-

tonicity condition that ensures equilibrium uniqueness. For short, write $\hat{\rho}^z_{\tilde{z}}$ for $q^{\rho_z}_{X|Z=\tilde{z}}$ and assume:

$$\hat{\rho}^{\ell}_{h}(H) - \hat{\rho}^{\ell}_{\ell}(H) < \rho_h(H) - \rho_\ell(H) < \hat{\rho}^{h}_{h}(H) - \hat{\rho}^{h}_{\ell}(H). \tag{13}$$

The condition compares how the switch from low to high effort increases the probability of high output H under the objective technology and under the principal's two subjective technologies. It states that this return to effort is larger, relative to the objective return, under the subjective technology formed after a high-effort sample, and smaller under the subjective technology formed after a low-effort sample.

Under the above conditions, the misspecified principal guided by the sample model chooses the same contract as the well-specified principal, but achieves a lower effort in the unique equilibrium.

Proposition 2. *When c_h is sufficiently low, the sample equilibrium is unique. The principal chooses the same wage profile, given in* (9)*, as the well-specified principal. The workers exert an interior effort: $0 < e^*(h) < 1$.*

The mechanism behind the result is the following. The principal's subjective assessment of the return to effort is shaped by the workers' behavior. As assumed in (13), when the true effort is high, outcomes appear especially diagnostic of effort in the principal's sample model, and thus the principal believes that a relatively low reward for H suffices to induce high effort. When the true effort is low, outcomes appear less diagnostic, and the principal believes that a higher reward is needed. Thus, none of the pure effort levels can appear in the sample equilibrium. The principal's perceived incentive-compatible reward is decreasing in the true effort level that generated the sample. This feedback creates an inverse demand for effort that intersects the workers' true effort supply curve at an interior effort level. See Figure 2B.[6]

To illustrate the proposition, consider the following parametrization. Let

$$r(L) = 0, \qquad r(N) = 1, \qquad r(H) = 2, \qquad c_h = 0.0184.$$

The objective technology and the principal's primitive theory are:

	objective technology ρ_z			primitive theory π_z		
	L	H	N	L	H	N
ℓ	0.240	0.180	0.580	0.350	0.150	0.500
h	0.216	0.364	0.420	0.150	0.350	0.500

;

[6]Condition (13) ensures both uniqueness and interiority of the equilibrium: it makes the inverse demand for effort decreasing and ensures that demand and supply intersect in the interior. Other parameter specifications may instead give rise to a unique pure equilibrium, with either high or low effort, or to three equilibria, one of which is mixed.

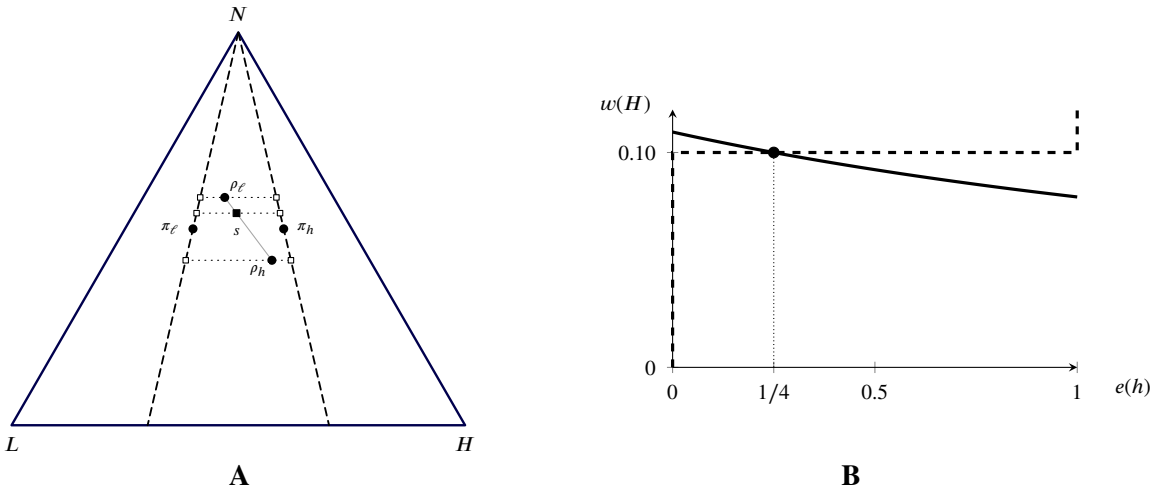


Figure 2: Display A: the objective and primitive types are shown as dots, the equilibrium sample s as a black square, and the principal's subjective types under various samples as white squares. Display B: the solid curve is the principal's inverse demand for effort and the dashed curve is the workers' supply correspondence.

see also Figure 2A. For this parametrization, all three principals reward only outcome H. The well-specified benchmark principal sets this reward to

$$w_{\text{b}}(H) = \frac{0.0184}{0.364 - 0.180} = 0.10.$$

The misspecified principal guided by the generative model instead chooses

$$w_{\text{g}}(H) = \frac{0.0184}{0.350 - 0.150} = 0.092,$$

and thus fails to induce high effort. In contrast, the principal guided by the sample model chooses the same contract as the well-specified counterpart. Yet, in the unique sample equilibrium, workers choose high effort with only interior probability $e^*(h) = \frac{1}{4}$.

Despite her misspecification, the principal guided by the sample model gets the equilibrium incentives just right: like her well-specified counterpart, she keeps the workers indifferent under the objective technology. Yet she still pays a price for misspecification. The well-specified principal could ensure full effort by increasing the reward for H slightly above the workers' true incentive-compatible level. The misspecified principal cannot do so in a stable way: full effort would generate a sample in which H appears especially diagnostic of high effort. The principal would then believe that a lower reward for H suffices, and would no longer want to maintain the high reward needed to sustain full effort.

This example echoes a broader theme in organizational economics, going back at least to Leibenstein (1966): formal contracts are not the only determinants of performance. Here, the well-specified and misspecified principals choose the same contract, yet the well-specified principal achieves a better outcome. The difference lies not in formal incentives, but in equilibrium feedback between behavior, observed data, and the principal's interpretation of the production technology. Misspecification can therefore lock the principal into worse outcomes even when the contract itself is unchanged.

5.3 Learning from Aggregate Behavior

Economic agents often infer the distribution of latent types in a population from aggregate behavior. Observing only the aggregate behavior of her peers, an investor may infer the distribution of market sentiment from their portfolios. Similarly, a firm may infer the distribution of consumer tastes from market shares. Such inference is delicate because the analyst uses the same aggregate sample both to infer the distribution of types and to interpret what each observed choice reveals about those types. An increased prevalence of risky investment, for example, may indicate more optimistic investors, but it may also lead the analyst to revise the interpretation of what risky investment indicates about sentiment.

Suppose the analyst is a retail investor who observes the fractions $s(x)$ of her peers choosing investment positions x; z encodes an investor's market sentiment. To guide her own investment choice, left unmodeled here, the analyst forms a belief q^s_Z about the prevalence of market sentiments in the peers' population. We decompose the response of the sample prior q_Z to a sample shift into direct and indirect effects. Assuming that a risky investment position stochastically indicates optimism, an observed increase in risky investment directly suggests a greater prevalence of optimism among her peers. The MLE-Bayes analyst, however, also adjusts her view of how sentiments are correlated with observed positions, which affects the adjustment of q_Z indirectly.

We fix a baseline sample s, and let the analyst form the baseline sample model q^s_{XZ}. We then consider a shift in the sample from s to s' and study the induced adjustment of the sample prior from q^s_Z to $q^{s'}_Z$. To isolate the direct effect, hold fixed the baseline conditional distributions $q^s_{Z|X=x}$ and define

$$b^{s'}_Z = \sum_{x \in \mathcal{X}} s'(x) q^s_{Z|X=x}.$$

This is the type distribution obtained by applying the baseline interpretation of choices to the new sample. The movement from q^s_Z to $b^{s'}_Z$ captures the direct effect of the sample shift. The wedge between $q^{s'}_Z$ and $b^{s'}_Z$ captures the indirect effect generated by the analyst's endogenous reinterpretation

of choices.

For any nonempty set $\mathcal{E} \subseteq \mathcal{Z}$ of types, let $\mathcal{S}(\mathcal{E}) := \{s : \operatorname{supp} q_Z^s = \mathcal{E}\}$ denote the set of samples for which the analyst employs exactly the types in $\mathcal{E}$. The sets $\mathcal{S}(\mathcal{E})$ admit a simple product structure that leads to an intuitive geometric characterization. Appendix B.3.1 presents the details. We can accordingly organize the comparative statics with respect to sample shifts into two components that together give a complete characterization. The first step, in Appendix B.3.1, considers how the sample shift affects the set $\mathcal{E}$ of latent types that appears in the analyst's sample prior. We focus here on the second component, considering local sample shifts within $\mathcal{S}(\mathcal{E})$ that do not affect the set of employed latent types.

To characterize fixed support cells $\mathcal{S}(\mathcal{E})$, we decompose each sample as a product of a fitted generative component and a discrepancy component. Let $\mathcal{P}(\mathcal{E})$ denote the relative interior of the convex hull of the employed primitive types $\{\pi_z\}_{z\in\mathcal{E}}$. Let $\mathcal{D}(\mathcal{E})$ be the set of discrepancy vectors $d \in \mathbb{R}^{\mathcal{X}}_{++}$ satisfying the optimality conditions

$$\pi_z \cdot d = 1 \text{ for all } z \in \mathcal{E}, \qquad \pi_z \cdot d \leq 1 \qquad \text{for all } z \notin \mathcal{E}.$$

For any $a, b \in \mathbb{R}^{\mathcal{X}}$, we define the coordinate-wise product $a \odot b = (a(x)b(x))_{x\in\mathcal{X}}$. Lemma 1 and rational expectations (Lemma 2(i)) imply the following corollary.

Corollary 1. *For each nonempty $\mathcal{E} \subseteq \mathcal{Z}$, the set $\mathcal{S}(\mathcal{E})$ of samples at which the analyst employs exactly the types in $\mathcal{E}$ is the coordinate-wise product of two polytopes:*

$$\mathcal{S}(\mathcal{E}) = \mathcal{P}(\mathcal{E}) \odot \mathcal{D}(\mathcal{E}) := \left\{p \odot d : p \in \mathcal{P}(\mathcal{E}), d \in \mathcal{D}(\mathcal{E})\right\}.$$

We identify two directions of sample shifts within a fixed support cell $\mathcal{S}(\mathcal{E})$ that have especially simple comparative statics. We call them the d-fiber and the p-fiber. Other local sample shifts within $\mathcal{S}(\mathcal{E})$ can be decomposed into two sequential movements, one along the d-fiber and the other along a p-fiber.

Write the baseline sample as $s = p_X \odot d$, where $p_X \in \mathcal{P}(\mathcal{E})$ is the fitted generative distribution and $d \in \mathcal{D}(\mathcal{E})$ is the discrepancy vector. The d-fiber,

$$\mathcal{F}_d(s) := \{p_X\} \odot \mathcal{D}(\mathcal{E}),$$

holds fixed the fitted generative distribution and varies only the discrepancy between this fitted distribution and the observed sample. Such shifts cannot be explained by changing only the prevalence

of the current sample types. The p-fiber,

$$\mathcal{F}_p(s) := \mathcal{P}(\mathcal{E}) \odot \{d\},$$

holds fixed the discrepancy and varies the fitted generative distribution. These shifts can be explained by changing only the prevalence of the current sample types, while holding their definitions fixed. See Figure 3.

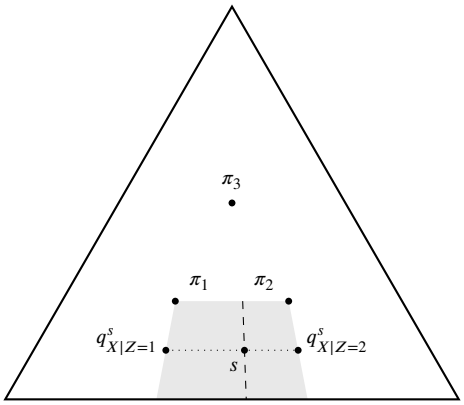


Figure 3: The grey area depicts the set $S(\{1,2\})$ of samples at which the analyst employs exactly types 1 and 2. The dashed segment is the d-fiber $\mathcal{F}_d(s)$, and the dotted segment is the p-fiber $\mathcal{F}_p(s)$.

The proposition below shows that the two fibers isolate two distinct comparative-static forces. Along the d-fiber, the MLE-Bayes analyst and the fixed-interpretation benchmark make the same inference: the inferred type distribution does not change. Along the p-fiber, the MLE-Bayes analyst reacts more sharply than the fixed-interpretation benchmark does.

Proposition 3. *1. If $s' \in \mathcal{F}_d(s)$, then the MLE-Bayes analyst and the benchmark maintain the initial type distribution:*

$$q^{s'}_Z = b^{s'}_Z = q^s_Z.$$

2. If $s' \in \mathcal{F}_p(s)$, then the MLE-Bayes analyst adjusts her prior more sharply than the benchmark: there exists a column-stochastic garbling matrix $M = (M_{z,z'})_{z,z' \in \mathcal{E}}$ such that $b^{s'}_Z = M q^{s'}_Z$.

For the first statement, consider a sample shift from s to s' within the d-fiber. The optimality conditions imply that

$$q^s_{X|Z=z} \cdot \frac{s'}{s} = 1 \quad \text{for all } z \in \mathcal{E}. \tag{14}$$

Thus, all sample types formed at the initial sample explain the sample shift within the d-fiber equally

well, so changing the types' prevalence does not explain the discrepancy s'/s. The sample prior then remains fixed and the MLE-Bayes and benchmark analysts' updating coincide.

By contrast, consider a sample shift from s to s' within the p-fiber. Such a shift lies in the convex hull of the baseline sample types $\{q^s_{X|Z=z}\}_{z\in\mathcal{E}}$. Hence, by Theorem 2, the MLE-Bayes analyst keeps these sample types fixed. She therefore attributes the sample shift entirely to a change in the prevalence of the same latent types. Since the benchmark partially absorbs the sample shift by adjusting the sample types, the MLE-Bayes adjustment of q_Z is sharper, as formalized in the proposition. Thus, when existing latent types can rationalize the shift, endogenous reinterpretation amplifies the inferred change in population composition.

6 Literature

We focus on the analyst's reasoning about a sample that does not match any parametrization of the analyst's statistical theory. Similar discrepancies between theories and data appear throughout the literature on misspecified learning, originating with Berk (1966) and White (1982) and extended to strategic interactions by Esponda and Pouzo (2016). This literature focuses on the analyst's estimate of the data-generating process. We focus instead on how the analyst represents the realized data themselves, forcing her to engage with the discrepancy between generative predictions and observations.[7]

We contribute to an emerging literature that studies how analysts revise their models after encountering surprising or poorly explained data. The analyst in Ortoleva (2012) reconsiders her prior after observing a sufficiently unlikely event. Cho and Kasa (2015) and Ba (2026) study model switching when the data are endogenously generated by the analyst's actions, while Gagnon-Bartsch et al. (2018) study whether endogenous attention exposes an analyst to evidence that discredits her model. Following the unawareness literature, Karni and Vierø (2013) consider an analyst who expands her state space after encountering an unforeseen observation. The analyst in Lanzani (2025) responds to evidence against her structured models by increasing her concern about misspecification. Our approach is distinctive in focusing on how the analyst revises her latent representation of the realized data themselves.

Our approach shares an information-geometry core with the standard misspecification literature. There, the asymptotic estimate minimizes the Kullback-Leibler divergence from the empirical sample frequencies. Similarly, our latent representation—a pair of the generative model p^s_{XZ} and the

[7] Frick et al. (2020, 2023, 2024) share our focus on misspecified learning about a latent state. While they focus on a fixed single state, we study learning in mixture models and, within this environment, how the analyst reasons about data inconsistent with her statistical model. See Bohren and Hauser (2025) for a survey of the misspecification literature.

sample model q^s_{XZ}—solves the information-geometric problem[8]

$$\begin{aligned} \min_{p_{XZ}, q_{XZ}} \ & \mathrm{KL}(q_{XZ} \| p_{XZ}) \\ \text{s.t.} \quad & p_{X|Z=z} = \pi_z \quad \text{for all } z \in \mathrm{supp}(p_Z), \\ & q_X = s. \end{aligned}$$

This geometric perspective is at the center of the Expectation-Maximization algorithm for latent variable models (Dempster et al., 1977) and of variational inference (Jordan et al., 1999; Blei et al., 2017). The simultaneous tracking of the generative model and the sample model also underlies the variational autoencoder of Kingma and Welling (2013) in machine learning. This connection motivates our terminological choice. We share the term "generative model" with the machine learning literature, while our sample model is the counterpart of the variational recognition model.[9] In contrast, the variational-inference literature treats the recognition model primarily as a computational device, introduced to approximate an intractable posterior and facilitate estimation of the generative model. Our paper instead treats the sample model as the object of economic interest and provides a novel axiomatic characterization of it.

Variational Bayes methods highlight a conservative aspect of Bayesian updating: the posterior can be characterized as the minimal departure from the prior subject to the informational constraint. A related minimum-distance approach is developed by Dominiak et al. (2023, 2025). Their analyst chooses, among the posterior beliefs consistent with the observed information, the belief closest to the prior. We share the view that updating can be understood as a conservative departure from prior beliefs, but study a different object: how latent type definitions are revised when the analyst represents the observed sample.

The two steps combined in the MLE-Bayes procedure are familiar in empirical work. Lazarsfeld and Henry (1968), Kiefer and Wolfowitz (1956), and Lindsay (1983) provide early and classic treatments of latent-type mixture models. Heckman and Singer (1984) illustrate how seemingly counterintuitive behavior can be explained using a mixture of latent types; see Bonhomme et al. (2022) for a recent analysis. Rather than expanding on practical empirical methods, we establish conservative properties of the MLE-Bayes procedure and discuss the procedure's economic implications.

We contribute to an emerging theoretical literature that departs from the standard Bayesian framework, modeling an economic agent as an intuitive statistician. Our analyst is an *empirical*

[8]For fixed p_{XZ}, the chain rule for relative entropy implies that the optimal q_{XZ} satisfies $q_{Z|X} = p_{Z|X}$; optimizing over p_Z then reduces to the maximum-likelihood problem.

[9]Aridor et al. (2020) and Aridor et al. (2025) apply variational autoencoders to economic decision-making.

Bayesian who uses the sample to estimate a "prior" belief and then updates according to Bayes' rule. Recent theoretical work includes Schwartzstein and Sunderam (2021) and Aina (2021). Our work differs in emphasizing the implications of misspecification and the analyst's reaction to it.

Although our economic focus is different, our sample-model problem is formally related to the rational-inattention model with entropic costs (Matějka and McKay, 2015). In this model, costly information acquisition generates a state-contingent stochastic choice rule. Our analyst instead starts from an observed sample. Yet, if observable outcomes x are formally reinterpreted as states, latent types z as actions, and log primitive likelihoods $\log \pi_z(x)$ as payoffs, the assignment rule $q^s_{Z|X=x}$ in our MLE-Bayes sample model coincides with the state-contingent stochastic choice rule in the corresponding entropic rational-inattention problem.[10]

This equivalence explains several structural parallels with the rational-inattention literature. Our diagnostic-conservatism condition is the analogue of the likelihood-ratio property of entropic rational-inattention rules (Caplin and Dean, 2015). Our local type-conservatism result in Theorem 2 is similarly related to local invariance properties of rational-inattention posteriors, and the set of employed types is analogous to the rationally inattentive consideration set in Caplin et al. (2019). These parallels reflect a common entropy geometry, but the economic objects are different: the rational-inattention model studies costly signal acquisition, whereas we study the latent representation of observed data.

7 Summary

This paper studies how an analyst applies a latent-type theory to a realized sample when the sample may not be representable as a mixture of the primitive types. The central object is a sample model: a joint belief over observable outcomes and latent types whose observable marginal matches the empirical distribution in the sample. When the primitive theory is misspecified, constructing such a model requires the analyst to adjust the theory's type definitions.

We characterize an intuitive MLE-Bayes procedure, in which the analyst first estimates a latent-type mixture model by maximum likelihood and then applies Bayes' rule to observations. The characterization shows that this familiar procedure is conservative. It adjusts the predictive content of the types so that the sample model matches the observed data, but it preserves their diagnostic content: the evidence that an observation provides in favor of one latent type over another remains tied to the primitive theory regardless of the observed evidence.

The distinction between the fitted generative model and the sample model complements Berk-

[10] We establish this equivalence in a previous draft of this project, Samuelson and Steiner (2025).

Nash reasoning. Berk-Nash reasoning identifies the best-fitting generative model within a misspecified theory. Our sample-model approach instead asks how the analyst represents the realized sample when the fitted generative model does not reproduce the empirical distribution of observables. The moral-hazard application shows that this distinction can matter for equilibrium behavior.

Finally, the MLE-Bayes procedure is also conservative across samples. Once the analyst has formed sample types, she preserves their definitions whenever the new sample can still be represented as a mixture of them. Type definitions are adjusted only when realism makes such adjustments necessary. Thus, when a change in the observed data can be explained by a change in the prevalence of existing sample types, the analyst attributes the shift to population composition rather than to a reinterpretation of the observations. This makes her inference about population composition especially responsive to such sample shifts.

A promising extension would allow the type definitions in the analyst's theory to interact with the analyst's actions. For example, the analyst may be a regulatory agency. Types might specify the return distributions generated by financial institutions. These distributions may depend on the institutions' due diligence and prudence in shaping their portfolios, which in turn may respond to the agency's regulatory actions.

A Proofs

A.1 Proof of Lemma 1

Proof. The maximum-likelihood optimization in (2), written in full detail, is:

$$\begin{aligned} \max_{(p_Z(z))_{z\in\mathcal{Z}}} \quad & \sum_{x\in\mathcal{X}} s(x)\,\ln p_X(x), \qquad \text{where} \qquad p_X(x)=\sum_{z\in\mathcal{Z}} p_Z(z)\pi_z(x) \\ \text{s.t.} \quad & \sum_{z\in\mathcal{Z}} p_Z(z)=1, \\ & p_Z(z)\geq 0 \text{ for all } z\in\mathcal{Z}. \end{aligned}$$

The objective is concave in p_Z and the feasible set $\Delta(\mathcal{Z})$ is compact and convex, so the Kuhn-Tucker conditions are necessary and sufficient. The Lagrangian is

$$\mathcal{L}(p_Z,\lambda,\mu)=\sum_{x\in\mathcal{X}} s(x)\ln p_X(x)+\lambda\Big(\sum_{z\in\mathcal{Z}} p_Z(z)-1\Big)+\sum_{z\in\mathcal{Z}}\mu_z\,p_Z(z),$$

with multipliers $\lambda \in \mathbb{R}$ and $\mu_z \geq 0$. The gradient of the objective is[11]

$$\frac{\partial}{\partial p_Z(z)} \sum_{x \in \mathcal{X}} s(x) \ln p_X(x) = \sum_{x \in \mathcal{X}} s(x) \frac{\pi_z(x)}{p_X(x)} = \pi_z \cdot d.$$

Thus, for each $z \in \mathcal{Z}$,

$$0 = \frac{\partial \mathcal{L}}{\partial p_Z(z)} = \pi_z \cdot d + \lambda + \mu_z.$$

Therefore,

$$\begin{aligned} \pi_z \cdot d &= -\lambda \text{ if } p_Z(z) > 0, \\ \pi_z \cdot d &\leq -\lambda \text{ if } p_Z(z) = 0. \end{aligned}$$

Multiplying $\pi_z \cdot d$ by $p_Z(z)$ and summing over z gives

$$\begin{aligned} \sum_{z \in \mathcal{Z}} p_Z(z) \left(\pi_z \cdot d \right) = \Big(\sum_{z \in \mathcal{Z}} p_Z(z) \pi_z \Big) \cdot d = p_X \cdot d \\ = \sum_{x \in \mathcal{X}} p_X(x) \frac{s(x)}{p_X(x)} = \sum_{x \in \mathcal{X}} s(x) = 1. \end{aligned}$$

But $\pi_z \cdot d = -\lambda$ for every $z \in \operatorname{supp} p_Z$, so the above identity implies $-\lambda = 1$, as needed. □

A.2 Proof of Lemma 2

Proof. Rational expectations: Recall from Lemma 1 that $d = s/p_X^s$ and that the estimate p_Z^s satisfies the optimality condition (4). By the construction of the sample model in (3),

$$q_{XZ}^{\mathsf{MLE},s}(x,z) = s(x) p_{Z|X=x}^s(z) = s(x) \frac{p_Z^s(z) \pi_z(x)}{p_X^s(x)} = p_Z^s(z) \pi_z(x) d(x).$$

Summing over x yields

$$q_Z^{\mathsf{MLE},s}(z) = p_Z^s(z)(\pi_z \cdot d).$$

Condition (4), $\pi_z \cdot d = 1$, implies $q_Z^{\mathsf{MLE},s}(z) = p_Z^s(z)$ for all $z \in \operatorname{supp} p_Z^s$ (and hence for all $z \in \mathcal{Z}$), as needed.

[11] The vector $d(x) := s(x)/p_X(x)$ is well-defined because the types are interior and hence $p_X(x) > 0$.

Diagnostic conservatism: Fix $z \in \operatorname{supp} q_Z^{\mathsf{MLE},s}$. Using $q_Z^{\mathsf{MLE},s} = p_Z^s$, we obtain

$$q_{X|Z=z}^{\mathsf{MLE},s}(x) = \frac{q_{XZ}^{\mathsf{MLE},s}(x,z)}{q_Z^{\mathsf{MLE},s}(z)} = \frac{p_Z^s(z)\pi_z(x)d(x)}{p_Z^s(z)} = \pi_z(x)d(x).$$

Applying this to $z, z' \in \operatorname{supp} q_Z^{\mathsf{MLE},s}$ and dividing yields A2. □

A.3 Proof of Proposition 1

Proof. Suppose there exists a sample model q_{XZ}^s that employs exactly the types in $\mathcal{E}$ and satisfies A1 and A2. Axiom A2 implies that there exists a tilt vector $d \in \mathbb{R}_{++}^{\mathcal{X}}$ such that

$$q_{X|Z=z}^s(x) = \pi_z(x)d(x) \quad \text{for all } x \in \mathcal{X} \text{ and all } z \in \operatorname{supp} q_Z^s.$$

Thus,

$$q_{XZ}^s(x,z) = q_Z^s(z)\pi_z(x)d(x). \tag{15}$$

Given q_Z^s, A1 uniquely identifies this tilt vector as

$$d(x) = \frac{s(x)}{\sum_{z\in\mathcal{Z}} q_Z^s(z)\pi_z(x)}. \tag{16}$$

Marginalizing (15) over x and dividing by $q_Z^s(z)$ yields

$$\pi_z \cdot d = 1 \qquad \text{for all } z \in \mathcal{E}. \tag{17}$$

Lemma 1, adapted to MLE supported on $\mathcal{E}$, implies that (17) is the first-order condition for the restricted maximum-likelihood problem on $\Delta(\mathcal{E})$. Since q_Z^s is interior in $\Delta(\mathcal{E})$ and the restricted likelihood is concave, q_Z^s is the restricted maximum-likelihood estimate:

$$q_Z^s = p_Z^{s,\mathcal{E}}. \tag{18}$$

The restricted estimate $p_Z^{s,\mathcal{E}}$ is unique by affine independence of the primitive types. Thus q_Z^s is unique. Given q_Z^s, (16) uniquely identifies d, and hence (15) uniquely identifies q_{XZ}^s.

It remains to prove that $q_{XZ}^s = q_{XZ}^{\mathsf{MLE},s,\mathcal{E}}$. Lemma 2 adapted to the MLE restricted to types in $\mathcal{E}$ implies that the restricted MLE-Bayes sample model satisfies rational expectations,

$$q_Z^{\mathsf{MLE},s,\mathcal{E}} = p_Z^{s,\mathcal{E}},$$

and diagnostic conservatism, so there exists a tilt vector $d^{\mathsf{MLE}} \in \mathbb{R}^{\mathcal{X}}_{++}$ such that

$$q_{XZ}^{\mathsf{MLE},s,\mathcal{E}}(x,z) = q_Z^{\mathsf{MLE},s,\mathcal{E}}(z)\pi_z(x)d^{\mathsf{MLE}}(x).$$

Since both q_{XZ}^s and $q_{XZ}^{\mathsf{MLE},s,\mathcal{E}}$ have observable marginal s, and since their Z-marginals coincide by (18) and rational expectations, equation (16) implies that their tilt vectors coincide:

$$d^{\mathsf{MLE}} = d.$$

Therefore,

$$q_{XZ}^{\mathsf{MLE},s,\mathcal{E}}(x,z) = p_Z^{s,\mathcal{E}}(z)\pi_z(x)d(x) = q_Z^s(z)\pi_z(x)d(x) = q_{XZ}^s(x,z).$$

□

A.4 Proof of Theorem 1

Proof. Only-if direction: Suppose that the map $s \mapsto q_{XZ}^s$ satisfies A1–A4. Fix $s \in \operatorname{int}\Delta(\mathcal{X})$, and let $(p_{XZ}^s, q_{XZ}^{\mathsf{MLE},s})$ be the unique MLE-Bayes latent representation associated with s. We need to show that

$$q_{XZ}^s = q_{XZ}^{\mathsf{MLE},s}.$$

First, consider the case $s \in \operatorname{co}(\mathcal{T})$. By A3, every employed sample type coincides with its corresponding primitive type. Together with A1 and affine independence, this implies that the sample prior q_Z^s expresses s as a unique mixture of the primitive types. The MLE prior p_Z^s is the same unique mixture decomposition of s, and hence

$$q_{XZ}^s(x,z) = q_Z^s(z)\pi_z(x) = p_Z^s(z)\pi_z(x) = q_{XZ}^{\mathsf{MLE},s}(x,z) \qquad \text{for all } (x,z),$$

as needed.

It remains to consider the case $s \notin \operatorname{co}(\mathcal{T})$. Let $\mathcal{E} := \operatorname{supp} p_Z^s$. It suffices to prove that

$$\operatorname{supp} q_Z^s = \mathcal{E}, \tag{19}$$

because then Proposition 1 implies that the axiomatic sample model is the unique MLE-Bayes sample model restricted to employing only types in $\mathcal{E}$,

$$q_{XZ}^s = q_{XZ}^{\mathsf{MLE},s,\mathcal{E}}.$$

Additionally, because the unrestricted MLE prior p_Z^s already has support $\mathcal{E}$, the restriction to $\mathcal{E}$ is not binding; hence

$$q_{XZ}^s = q_{XZ}^{\mathsf{MLE},s,\mathcal{E}} = q_{XZ}^{\mathsf{MLE},s}.$$

In fact, instead of (19), we prove a stronger statement,

$$q_Z^s = p_Z^s, \tag{20}$$

which implies (19). The stronger statement (20) is useful because we can prove it by connecting the sample s to the fitted observable distribution p_X^s, using the continuity axiom A4. For this, we define a path of samples

$$s_t := ts + (1-t)p_X^s, \qquad t \in [0,1],$$

and note that $s_t \in \operatorname{int} \Delta(\mathcal{X})$ for all such t. The role of this path is to start from a perfect-fit sample at $t = 0$, where A1 and A3 identify the prior, and then to use A4 to rule out any change in the support of the axiomatic sample model before reaching $t = 1$.

We first establish that the MLE prior is constant along this path:

$$p_Z^{s_t} = p_Z^s \qquad \text{for all } t \in [0,1]. \tag{21}$$

To show this, fix t and consider p_Z^s as a candidate prior for the MLE problem at sample s_t. The X-marginal induced by this candidate is p_X^s. Hence the associated discrepancy vector is

$$d_t := \frac{s_t}{p_X^s} = t\frac{s}{p_X^s} + (1-t)\mathbf{1} = td_1 + (1-t)\mathbf{1}, \qquad \text{where } d_1 := \frac{s}{p_X^s}.$$

The optimality conditions in Lemma 1 hold for p_Z^s and d_1. Because each π_z is a well-normalized distribution, $\pi_z \cdot \mathbf{1} = 1$ for every z. Thus the same conditions also hold for p_Z^s and d_t:

$$\begin{aligned}
\pi_z \cdot d_t &= t(\pi_z \cdot d_1) + (1-t)(\pi_z \cdot \mathbf{1}) = 1 \quad \text{for every } z \in \mathcal{E},\\
\pi_z \cdot d_t &= t(\pi_z \cdot d_1) + (1-t)(\pi_z \cdot \mathbf{1}) \le 1 \quad \text{for every } z \notin \mathcal{E}.
\end{aligned}$$

Using uniqueness of the MLE prior, Lemma 1 therefore implies $p_Z^{s_t} = p_Z^s$. This proves (21).

Now define a set

$$C := \left\{ t \in [0,1] : q_Z^{s_t} = p_Z^s \right\}.$$

We have $0 \in C$, because $s_0 = p_X^s \in \operatorname{co}(\mathcal{T})$, the perfect-fit argument from the first case gives $q_Z^{s_0} = p_Z^{s_0}$, and (21) gives $p_Z^{s_0} = p_Z^s$.

We use continuity (A4) for an expansion argument that will establish that $1 \in C$ too. By A4, the map $t \mapsto q_Z^{s_t}$ is continuous, so C is closed in $[0,1]$. We show below that C is also relatively open in $[0,1]$, which completes the proof of the *only-if* direction: The set C is nonempty because $0 \in C$, closed in $[0,1]$, and relatively open in $[0,1]$. It follows that $C = [0,1]$. In particular, $1 \in C$. Since $s_1 = s$, this gives (20). As argued above, (20) implies $q_{XZ}^{s} = q_{XZ}^{\mathsf{MLE},s}$.

It remains to prove that C is relatively open. For this, fix $t^* \in C$, and write $s^* := s_{t^*}$. Since $t^* \in C$,

$$q_Z^{s^*} = p_Z^{s},$$

and hence $\operatorname{supp} q_Z^{s^*} = \mathcal{E}$. Since the map $t \mapsto q_Z^{s_t}$ is continuous, we have $q_Z^{s_t}(z) > 0$ for every $z \in \mathcal{E}$ and for all t sufficiently close to t^*. Thus, there is a relative neighborhood U of t^* in $[0,1]$ such that

$$\mathcal{E} \subseteq \operatorname{supp} q_Z^{s_t} \qquad \text{for all } t \in U. \tag{22}$$

We next show that the support cannot strictly contain $\mathcal{E}$. Suppose, for a contradiction, that for some $t \in [0,1]$,

$$\mathcal{F} := \operatorname{supp} q_Z^{s_t} \supsetneq \mathcal{E}.$$

By Proposition 1, the sample model $q_{XZ}^{s_t}$ must coincide with the MLE-Bayes sample model restricted to employing only types in $\mathcal{F}$. Hence, specifically for the Z-marginals,

$$q_Z^{s_t} = q_Z^{\mathsf{MLE},s_t,\mathcal{F}}.$$

But by (21), the unrestricted MLE prior at s_t is $p_Z^{s_t} = p_Z^{s}$, whose support is $\mathcal{E}$. Since $\mathcal{E} \subsetneq \mathcal{F}$, this unrestricted maximizer is feasible in the problem restricted to $\mathcal{F}$. Hence the restriction to $\mathcal{F}$ is not binding, and the uniqueness of the MLE prior, together with rational expectations, implies

$$q_Z^{\mathsf{MLE},s_t,\mathcal{F}} = q_Z^{\mathsf{MLE},s_t} = p_Z^{s_t} = p_Z^{s}.$$

Therefore $\operatorname{supp} q_Z^{s_t} = \mathcal{E}$, contradicting $\operatorname{supp} q_Z^{s_t} = \mathcal{F} \supsetneq \mathcal{E}$.

Combining the established observation that the support cannot expand with (22), we obtain

$$\operatorname{supp} q_Z^{s_t} = \mathcal{E} \qquad \text{for all } t \in U.$$

Applying Proposition 1, or equivalently its Z-marginal implication, yields

$$q_Z^{s_t} = p_Z^{s_t,\mathcal{E}}.$$

Since $p^{s_t}_Z = p^s_Z$ by (21) and this prior has support $\mathcal{E}$, the restriction to $\mathcal{E}$ is not binding. Hence

$$q^{s_t}_Z = p^{s_t,\mathcal{E}}_Z = p^{s_t}_Z = p^s_Z.$$

Thus C is relatively open. This completes the proof of the *only-if* direction.

If direction: We verify that $q^{\mathsf{MLE},s}_{XZ}$ satisfies A1–A4. A1 follows immediately from the construction of $q^{\mathsf{MLE},s}_{XZ}$ in (3). A2 was established in Lemma 2(ii). A3 is also immediate because when $s \in \mathrm{co}(\mathcal{T})$, then the generative model p^s_{XZ} perfectly matches the sample, $p^s_X = s$, and hence $q^{\mathsf{MLE},s}_{XZ} = p^s_{XZ}$. Then, $q^{\mathsf{MLE},s}_{X|Z=z} = p^s_{X|Z=z} = \pi_z$ for all $z \in \operatorname{supp} q^{\mathsf{MLE},s}_Z$, as needed for A3.

A4 follows from continuity of p^s_Z, because $q^{\mathsf{MLE},s}_Z = p^s_Z$ by Lemma 2(i). To see that p^s_Z is continuous in s, note that the likelihood objective is continuous in (s, p_Z), and the feasible set $\Delta(\mathcal{Z})$ is compact and independent of s. The Maximum Theorem therefore implies that the argmax correspondence is upper hemicontinuous. Since affine independence makes the maximizer unique for every s, this correspondence is single-valued and hence continuous. Thus $s \mapsto q^{\mathsf{MLE},s}_Z$ is continuous, proving A4. □

A.5 Proof of Theorem 2

Proof. By Theorem 1, it suffices to prove the claim for the MLE-Bayes sample model $q^s_{XZ} = q^{\mathsf{MLE},s}_{XZ}$.

Fix a sample s. Let $\mathcal{E} := \operatorname{supp} q^s_Z$ and $d := s/p^s_X$. By rational expectations (Lemma 2(i)), $q^s_Z = p^s_Z$. Hence, for every $z \in \mathcal{E}$,

$$q^s_{X|Z=z}(x) = \frac{q^s_{XZ}(x,z)}{q^s_Z(z)} = \frac{s(x)p^s_{Z|X=x}(z)}{p^s_Z(z)} = \frac{s(x)p^s_Z(z)\pi_z(x)/p^s_X(x)}{p^s_Z(z)} = \pi_z(x)d(x).$$

By Lemma 1,

$$\pi_z \cdot d = 1 \quad \text{for all } z \in \mathcal{E}, \qquad \pi_z \cdot d \le 1 \quad \text{for all } z \notin \mathcal{E}.$$

We now prove the two directions. Realism (A1) implies 1. ⇒ 2. For the converse, suppose $s' \in \mathrm{co}\big(\{q^s_{X|Z=z}\}_{z\in\mathcal{E}}\big)$ so that there exists $\mu \in \Delta(\mathcal{E})$ such that

$$s' = \sum_{z\in\mathcal{E}} \mu(z)\, q^s_{X|Z=z} = \sum_{z\in\mathcal{E}} \mu(z)\,(\pi_z \odot d) = \Big(\sum_{z\in\mathcal{E}} \mu(z)\pi_z\Big) \odot d.$$

Let

$$\tilde{p}_X := \sum_{z\in\mathcal{E}} \mu(z)\pi_z.$$

Then $d = s'/\tilde{p}_X$ is the discrepancy vector associated with the prior μ at sample s'.

Since $\operatorname{supp}\mu \subseteq \mathcal{E}$, we have

$$\pi_z \cdot d = 1 \quad \text{for all } z \in \operatorname{supp}\mu, \qquad \pi_z \cdot d \leq 1 \quad \text{for all } z \notin \operatorname{supp}\mu.$$

Therefore μ satisfies the optimality conditions in Lemma 1 for sample s'. Hence μ is the MLE prior at s', i.e. $p^{s'}_Z = \mu$.

By Lemma 2(i), $q^{s'}_Z = p^{s'}_Z = \mu$, and thus, for every $z \in \operatorname{supp} q^{s'}_Z$,

$$q^{s'}_{X|Z=z}(x) = \frac{s'(x)p^{s'}_{Z|X=x}(z)}{q^{s'}_Z(z)} = \frac{s'(x)p^{s'}_Z(z)\pi_z(x)/p^{s'}_X(x)}{p^{s'}_Z(z)} = \pi_z(x)d(x) = q^{s}_{X|Z=z}(x).$$

□

B Computations and Proofs for the Applications

B.1 Stereotypes, Section 5.1

We first use Lemma 1 to identify the MLE generative model. Consider the candidate generative prior

$$p^s_Z(\ell) = p^s_Z(h) = \frac{1}{2}.$$

It induces the generative prediction

$$p^s_X = \frac{1}{2}\pi_\ell + \frac{1}{2}\pi_h = \left(\frac{9}{20}, \frac{3}{20}, \frac{1}{10}, \frac{3}{10}\right),$$

and hence the discrepancy vector

$$d := \frac{s}{p^s_X} = \left(\frac{5}{9}, \frac{5}{3}, \frac{5}{2}, \frac{5}{6}\right). \tag{23}$$

The candidate satisfies the optimality conditions of Lemma 1, since both types are employed and

$$\pi_\ell \cdot d = \frac{1}{3} + \frac{1}{3} + \frac{1}{4} + \frac{1}{12} = 1, \qquad \pi_h \cdot d = \frac{1}{6} + \frac{1}{6} + \frac{1}{4} + \frac{5}{12} = 1.$$

Since

$$q^s_{X|Z=z}(x) = \pi_z(x)d(x),$$

for each z, combining (23) with the assumed primitive types π_z yields the sample types

$$q^s_{X|Z=\ell} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}), \qquad q^s_{X|Z=h} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{4}, \frac{5}{12}).$$

B.2 Moral Hazard, Section 5.2

Lemma 3. *If* (11) *holds, then the sample model employs both types at every sample*

$$s_\alpha = (1-\alpha)\rho_\ell + \alpha\rho_h, \qquad \alpha \in [0,1].$$

Proof. Fix a sample $s = s_\alpha$ and denote

$$a := \frac{\pi_\ell(H)}{\pi_\ell(L)}, \qquad b := \frac{\pi_h(H)}{\pi_h(L)}, \qquad \sigma := s(N).$$

Assumption (11) implies

$$a < \frac{s_\alpha(H)}{s_\alpha(L)} < b.$$

Consider the following two candidate sample types, written in the order (L, H, N):

$$\bar{q}^s_{X|Z=\ell} = \left(\frac{1-\sigma}{1+a}, \frac{a(1-\sigma)}{1+a}, \sigma\right), \qquad \bar{q}^s_{X|Z=h} = \left(\frac{1-\sigma}{1+b}, \frac{b(1-\sigma)}{1+b}, \sigma\right).$$

We verify that these are indeed the MLE-Bayes sample types.

These two candidate types satisfy A2. Moreover, since $s(H)/s(L) \in (a,b)$ and all three distributions assign the same probability σ to N, the sample s lies in the relative interior of the line segment connecting $\bar{q}^s_{X|Z=\ell}$ and $\bar{q}^s_{X|Z=h}$. Hence there exists $\mu \in (0,1)$ such that

$$s = (1-\mu)\bar{q}^s_{X|Z=\ell} + \mu\bar{q}^s_{X|Z=h}.$$

It remains to verify that these candidate sample types arise from the MLE-Bayes procedure. The two candidates have a common tilt relative to the primitive types: for some $d \in \mathbb{R}^{\mathcal{X}}_{++}$,

$$\bar{q}^s_{X|Z=z}(x) = \pi_z(x)d(x), \qquad z \in \{\ell, h\}.$$

Because both candidates are probability distributions, their normalization implies

$$\pi_\ell \cdot d = \pi_h \cdot d = 1.$$

Moreover, by construction $s = (1-\mu)\bar{q}^s_{X|Z=\ell} + \mu\bar{q}^s_{X|Z=h}$, and hence

$$s = \big((1-\mu)\pi_\ell + \mu\pi_h\big) \odot d.$$

Thus $d = s/p_X$ for the generative marginal induced by the prior $p_Z(\ell) = 1-\mu$, $p_Z(h) = \mu$. The equalities $\pi_\ell \cdot d = \pi_h \cdot d = 1$ are exactly the MLE optimality conditions. Hence this prior is the MLE prior. Since $0 < \mu < 1$, both types are employed, and the MLE-Bayes sample types are precisely the two candidate types above. □

Proof of Proposition 2. Since H maximizes the likelihood ratio $\pi_h(x)/\pi_\ell(x)$ across x, and the sample types preserve the likelihood ratio by A2, the expenditure-minimizing wage profile that implements high effort under the subjective technology $\hat{\rho}^s$ rewards only H.

We now characterize the expenditure-minimizing wage $w(H)$ chosen by the principal who weakly implements high effort. Let the observed sample be $s_\alpha = (1-\alpha)\rho_\ell + \alpha\rho_h$, $\alpha = e(h)$. Write $q^\alpha_z := q^{s_\alpha}_{X|Z=z}$, $z \in \{\ell, h\}$. Let

$$\lambda_H := \frac{\pi_h(H)}{\pi_\ell(H)} > 1, \qquad \lambda_L := \frac{\pi_h(L)}{\pi_\ell(L)} < 1.$$

By A2, the sample likelihood ratios satisfy

$$\frac{q^\alpha_h(H)}{q^\alpha_\ell(H)} = \lambda_H, \qquad \frac{q^\alpha_h(L)}{q^\alpha_\ell(L)} = \lambda_L, \qquad \frac{q^\alpha_h(N)}{q^\alpha_\ell(N)} = 1.$$

Hence $q^\alpha_h(N) = q^\alpha_\ell(N) = s_\alpha(N)$, where we used realism (A1) for the last equality. Combining A2, realism at N, and the normalization of q^α_ℓ and q^α_h gives

$$q^\alpha_\ell(H) = \frac{1-\lambda_L}{\lambda_H - \lambda_L}\big(1 - s_\alpha(N)\big),$$

and

$$q^\alpha_h(H) = \frac{\lambda_H(1-\lambda_L)}{\lambda_H - \lambda_L}\big(1 - s_\alpha(N)\big).$$

Therefore the subjective return to high effort is

$$\Delta^\alpha := q^\alpha_h(H) - q^\alpha_\ell(H) = \frac{(\lambda_H - 1)(1-\lambda_L)}{\lambda_H - \lambda_L}\big(1 - s_\alpha(N)\big). \tag{24}$$

The binding incentive-compatibility constraint yields the principal's inverse demand for effort:

$$W(\alpha) = \frac{c_h}{\Delta^\alpha}.$$

At the endpoints,

$$\Delta^0 = \hat{\rho}_h^\ell(H) - \hat{\rho}_\ell^\ell(H), \qquad \Delta^1 = \hat{\rho}_h^h(H) - \hat{\rho}_\ell^h(H).$$

By the condition in (13),

$$\Delta^0 < \rho_h(H) - \rho_\ell(H) < \Delta^1.$$

Since $s_\alpha(N)$ is affine in α, (24) implies that Δ^α is affine in α. Moreover, the condition in (13) gives $\Delta^0 < \Delta^1$. Hence Δ^α is strictly increasing in α, and therefore $W(\alpha)$ is strictly decreasing.

For c_h sufficiently small, the principal indeed wants to implement high effort. Under the subjective technology $\hat{\rho}^{s_\alpha}$, the gross revenue gain from high rather than low effort is positive. Indeed, $q_h^\alpha(N) = q_\ell^\alpha(N)$, while high effort shifts probability mass from L to H, so

$$\sum_x r(x)\big(q_h^\alpha(x) - q_\ell^\alpha(x)\big) = \Delta^\alpha\big(r(H) - r(L)\big) > 0.$$

The expected wage expenditure associated with the minimum wage that weakly implements high effort vanishes as $c_h \to 0$. Thus, for c_h sufficiently small, implementing high effort is strictly optimal for the principal.

The workers' reservation wage for high effort is

$$w_{\mathrm{b}}(H) = \frac{c_h}{\rho_h(H) - \rho_\ell(H)}.$$

The endpoint condition implies

$$W(0) > w_{\mathrm{b}}(H) > W(1).$$

Since W is continuous and strictly decreasing, there is a unique $\alpha^* \in (0,1)$ such that $W(\alpha^*) = w_{\mathrm{b}}(H)$.

At the sample s_{α^*}, the principal therefore chooses

$$w^*(H) = w_{\mathrm{b}}(H), \qquad w^*(L) = w^*(N) = 0.$$

Given this wage profile, workers are indifferent between low and high effort under the objective technology, so the mixed effort $e^*(h) = \alpha^*$ is optimal. The sample generated by this effort is $s^* = s_{\alpha^*}$, so (e^*, s^*, w^*) is a sample equilibrium.

The sample equilibrium is unique: If $\alpha = 0$, then $W(0) > w_b(H)$, so workers strictly prefer high effort. Hence $\alpha = 0$ cannot be an equilibrium. If $\alpha = 1$, then $W(1) < w_b(H)$, so workers strictly prefer low effort. Hence $\alpha = 1$ cannot be an equilibrium. Any interior equilibrium must satisfy $W(\alpha) = w_b(H)$, and this equation has the unique solution α^*. □

B.3 Learning from Aggregate Behavior, Section 5.3

B.3.1 Employed Types

Recall from Corollary 1 that the set $S(\mathcal{E})$ of samples at which the analyst employs exactly the types in $\mathcal{E}$ decomposes as $S(\mathcal{E}) = \mathcal{P}(\mathcal{E}) \odot \mathcal{D}(\mathcal{E})$, where $\mathcal{P}(\mathcal{E})$ is the relative interior of the convex hull of $\{\pi_z\}_{z\in\mathcal{E}}$ and $\mathcal{D}(\mathcal{E}) \subset \mathbb{R}^{\mathcal{X}}_{++}$ is the set of discrepancy vectors that satisfy the optimality conditions from Lemma 1 with the types in $\mathcal{E}$ employed.

The sets $\mathcal{D}(\mathcal{E})$ admit a geometric characterization, illustrated in Figure 4.[12] For each type $z \in \mathcal{Z}$, consider the hyperplane

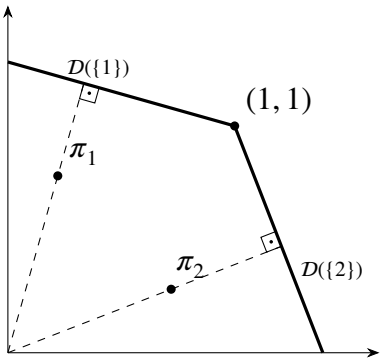
$$H_z = \left\{ d \in \mathbb{R}^{\mathcal{X}} : \pi_z \cdot d = 1 \right\},$$

which is orthogonal to the ray spanned by π_z and passes through $(1, 1, \ldots, 1)$. Let

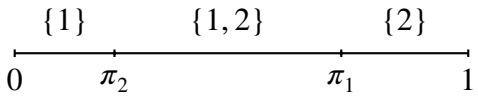
$$\tilde{\mathcal{D}} = \left\{ d \in \mathbb{R}^{\mathcal{X}}_{++} : \pi_z \cdot d \leq 1 \text{ for all } z \in \mathcal{Z} \right\}$$

denote the intersection of the corresponding half-spaces, restricted to the positive orthant. Then, for each set $\mathcal{E}$, the set $\mathcal{D}(\mathcal{E})$ is the face of $\tilde{\mathcal{D}}$ supported by the hyperplanes $\{H_z\}_{z\in\mathcal{E}}$.

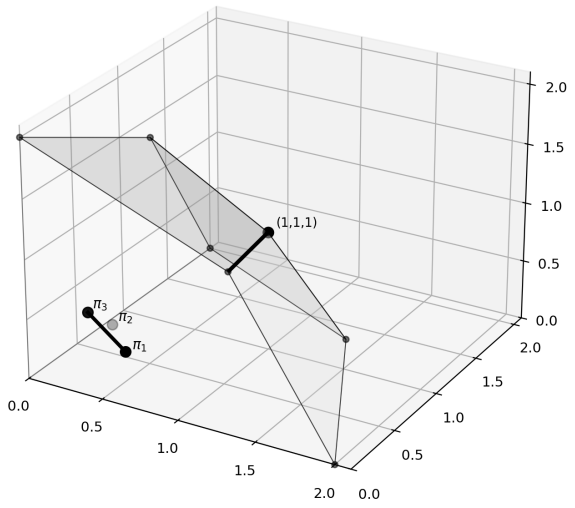
[12] Although the sets $S(\mathcal{E})$ depicted in Figure 4 are polytopes, this feature does not extend to higher dimensions, because coordinate-wise products of two polytopes need not have linear boundaries.



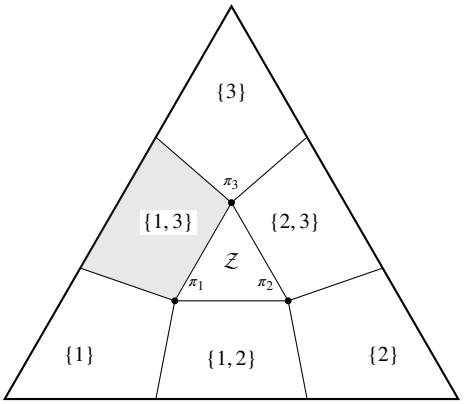
A



B



C



D

Figure 4: Displays A and B: binary setting with $|\mathcal{X}| = 2$ and primitive types π_1 and π_2. Display A represents both probability distributions and discrepancy vectors d as points in $\mathbb{R}^2_+$. For each singleton set $\mathcal{E} = \{z\}$, $z = 1, 2$, the set of feasible discrepancy vectors $\mathcal{D}(\mathcal{E})$ is the bold segment originating at $(1, 1)$ and perpendicular to the ray spanned by π_z. When both primitive types are employed, $\mathcal{E} = \{1, 2\}$, the set $\mathcal{D}(\mathcal{E})$ reduces to the singleton set $\{(1, 1)\}$. Display B depicts the binary simplex as the interval $[0, 1]$ and partitions it into three cells, $\mathcal{S}(\mathcal{E})$ for $\mathcal{E} = \{1\}, \{1, 2\}, \{2\}$. Each cell $\mathcal{S}(\mathcal{E}) = \mathcal{P}(\mathcal{E}) \odot \mathcal{D}(\mathcal{E})$ is the coordinate-wise product of the convex hull $\mathcal{P}(\mathcal{E})$ of the primitive types in $\mathcal{E}$ and the corresponding set of discrepancy vectors $\mathcal{D}(\mathcal{E})$. For example, when $\mathcal{E} = \{1, 2\}$, we have $\mathcal{D}(\mathcal{E}) = \{(1, 1)\}$, so no discrepancy arises between the generative prediction p^s_X and the sample s. Accordingly, $\mathcal{S}(\{1, 2\})$ coincides with the convex hull of the two primitive types.

Displays C and D are analogous for $|\mathcal{X}| = 3$ and three primitive types, $\mathcal{Z} = \{1, 2, 3\}$. The highlighted segments in C are $\mathcal{P}(\{1, 3\})$—the convex hull of these two primitive types—and $\mathcal{D}(\{1, 3\})$—the set of discrepancy vectors that can occur when these two types are employed. The highlighted cell in display D is the set of samples for which this pair of types is optimally employed: $\mathcal{S}(\{1, 3\}) = \mathcal{P}(\{1, 3\}) \odot \mathcal{D}(\{1, 3\})$.

B.3.2 Proof of Proposition 3

Proof. *Statement 1* follows from the fact that if $s' \in \mathcal{F}_d(s)$, then, by Lemma 1, p^s_Z satisfies the optimality conditions for the sample s'; hence $p^{s'}_Z = p^s_Z$. Then, by rational expectations (Lemma 2(i)), $q^s_Z = q^{s'}_Z$. Thus, by the construction of the benchmark, the MLE-Bayes analyst and the benchmark infer the same type distribution.

Statement 2: Since, by Theorem 2,

$$q^{s'}_{X|Z=z} = q^s_{X|Z=z}, \qquad \text{for all } z \in \mathcal{E},$$

we can write

$$s'(x) = \sum_{z\in\mathcal{E}} q^{s'}_Z(z)\, q^s_{X|Z=z}(x).$$

The benchmark prior at s' is

$$b^{s'}_Z(z) = \sum_{x\in\mathcal{X}} b^{s'}_{XZ}(x,z) = \sum_{x\in\mathcal{X}} s'(x)\, p^s_{Z|X=x}(z).$$

Substituting the decomposition of s' gives

$$b^{s'}_Z(z) = \sum_{z'\in\mathcal{E}} q^{s'}_Z(z') \sum_{x\in\mathcal{X}} q^s_{X|Z=z'}(x)\, p^s_{Z|X=x}(z).$$

Define, for $z, z' \in \mathcal{E}$,

$$M_{z,z'} := \sum_{x\in\mathcal{X}} q^s_{X|Z=z'}(x)\, p^s_{Z|X=x}(z).$$

Then,

$$b^{s'}_Z = M q^{s'}_Z.$$

Finally, we verify that M is a garbling matrix. For each $z' \in \mathcal{E}$,

$$\sum_{z\in\mathcal{E}} M_{z,z'} = \sum_{x\in\mathcal{X}} q^s_{X|Z=z'}(x) \sum_{z\in\mathcal{E}} p^s_{Z|X=x}(z) = \sum_{x\in\mathcal{X}} q^s_{X|Z=z'}(x) = 1.$$

Thus, M is column-stochastic, and therefore $b^{s'}_Z$ is a garbling of $q^{s'}_Z$. □

References

Aina, C. (2021). Tailored stories. Technical report, Mimeo.

Aridor, G., R. A. da Silveira, and M. Woodford (2025). Information-constrained coordination of economic behavior. *Journal of Economic Dynamics and Control 172*, 104985.

Aridor, G., F. Grechi, and M. Woodford (2020). Adaptive efficient coding: A variational autoencoder approach. biorxiv preprint 2020–05, Cold Spring Harbor Laboratory.

Ba, C. (2026). Robust misspecified models. *American Economic Review 116*(4), 1340–1379.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics 37*(1), 51–58.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Bohren, J. A. and D. N. Hauser (2025). Misspecified models in learning and games. *Annual Review of Economics 17*, 427–451.

Bonhomme, S., T. Lamadon, and E. Manresa (2022). Discretizing unobserved heterogeneity. *Econometrica 90*(2), 625–643.

Caplin, A. and M. Dean (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review 105*(7), 2183–2203.

Caplin, A., M. Dean, and J. Leahy (2019). Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies 86*(3), 1061–1094.

Cho, I.-K. and K. Kasa (2015). Learning and model validation. *The Review of Economic Studies 82*(1), 45–82.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (methodological) 39*(1), 1–22.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2023). Inertial updating. arXiv preprint arXiv:2303.06336.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2025). Inertial updating with general information. arXiv preprint arXiv:2502.00958.

Esponda, I. and D. Pouzo (2016). Berk–Nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica 84*(3), 1093–1130.

Frick, M., R. Iijima, and Y. Ishii (2020). Misinterpreting others and the fragility of social learning. *Econometrica 88*(6), 2281–2328.

Frick, M., R. Iijima, and Y. Ishii (2023). Belief convergence under misspecified learning: A martingale approach. *The Review of Economic Studies 90*(2), 781–814.

Frick, M., R. Iijima, and Y. Ishii (2024). Welfare comparisons for biased learning. *American Economic Review 114*(6), 1612–1649.

Gagnon-Bartsch, T., M. Rabin, and J. Schwartzstein (2018). *Channeled attention and stable errors*. Harvard Business School Boston.

Heckman, J. and B. Singer (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*(2), 271–320.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine learning 37*, 183–233.

Karni, E. and M.-L. Vierø (2013). Reverse bayesianism: a choice-based theory of growing awareness. *American Economic Review 103*(7), 2790–2810.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational Bayes. arxiv preprint arxiv:1312.6114, Cornell University.

Lanzani, G. (2025). Dynamic concern for misspecification. *Econometrica 93*(4), 1333–1370.

Lazarsfeld, P. F. and N. W. Henry (1968). *Latent Structure Analysis*. Houghton Mifflin.

Leibenstein, H. (1966). Allocative efficiency vs." x-efficiency". *The American economic review 56*(3), 392–415.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *The annals of statistics 11*(1), 86–94.

Matějka, F. and A. McKay (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review 105*(1), 272–298.

Ortoleva, P. (2012). Modeling the change of paradigm: Non-Bayesian reactions to unexpected news. *American Economic Review 102*(6), 2410–2436.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 157–163. University of California Press.

Samuelson, L. and J. Steiner (2025). Robust latent data representations.

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

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.