Samuelson, Larry; Steiner, Jakub

Working Paper

Robust latent data representations

Working Paper, No. 460

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Samuelson, Larry; Steiner, Jakub (2025) : Robust latent data representations, Working Paper, No. 460, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-264855

This Version is available at:
https://hdl.handle.net/10419/339345

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 460

Robust Latent Data Representations

Larry Samuelson and Jakub Steiner

Revised version, July 2025

Robust Latent Data Representations*

Larry Samuelson
Yale University

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

July 21, 2025

Abstract

Economic agents often infer latent structures—such as preference types—from data, without exogenously specified priors. We model such agents as empirical Bayesians. They estimate both the prior over types and the meanings of types via maximum likelihood. We show this estimation is equivalent to decomposing the sample into subsamples, each best explained by a single available latent type, with the decomposition minimizing the average misfit. The equivalence yields structural properties: optimal latent representations are robust (type definitions locally invariant to data changes) and simple (type count bounded). We extend these properties to agents who face frictions in evaluating likelihoods.

1 Introduction

Economic agents frequently infer underlying latent structures from observations. From firms inferring preferences based on consumer choices to investors deducing private information from others' market behavior, economic agents act as analysts, drawing inferences about unobserved variables from observed outcomes. How can such latent structures be inferred in the absence of an exogenously specified prior?

*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.

This paper studies the economics of constructing latent representations. We consider an analyst—perhaps an empirical economist, a firm poring over consumer data, or a lay person making sense of their environment—who observes a sample of outcomes. The analyst lacks an exogenous specification of the meaning and prior distribution of the latent types generating these outcomes. She accordingly adopts an empirical Bayes approach: she estimates, via maximum likelihood, the distribution over latent types that best explains the observed sample and then proceeds as a Bayesian. The estimated structure—comprising the inferred prior and type-specific data-generating processes—constitutes a *mixture estimate*. Examples of mixture estimates abound, such as a seller estimating a mixed logit model from customer purchase histories to infer taste distributions; a human resources manager categorizing employees into discrete types based on multidimensional traits; or machine learning algorithms structuring complex data via mixtures of simpler generative processes.

The selection of a mixture estimate constitutes the choice of an information structure—a joint distribution over latent and observed variables—but the nature of the problem differs from standard information design settings, such as Bayesian persuasion or rational inattention. In the standard formulation, the agent holds a fixed prior over the latent variable and selects a correlation structure that determines the marginal distribution of observables. By contrast, our agent similarly selects a correlation structure, but with the marginal distribution over observables fixed to match empirical frequencies, while the distribution over the latent variable—an unobservable construct in this context—is freely adjusted.

Our central contribution is to establish that this empirical Bayes procedure is mathematically equivalent to a *sample-decomposition* problem. To estimate the distribution over latent types, the analyst decomposes the observed sample (e.g., the customer base) into distinct subsamples (e.g., market segments), each optimally associated with a single, pure latent type from a specified set of possible types. This decomposition minimizes the average lack of fit—measured by the Kullback-Leibler divergence—between each subsample's empirical distribution and the best-fitting data-generating process for that subsample. The estimated probability of each latent type corresponds to the associated subsample's share of the total data in the optimal decomposition. We refer to the pair of the sample decomposition and the resulting mixture estimate as a *latent representation*.

This equivalence reveals a structure familiar from the information design

literature. Just as a sender in a persuasion game optimally decomposes a prior into a distribution over posteriors to suit her objective, our analyst decomposes her sample of observations into a collection of subsamples so as to minimize average modeling error. This connection enables the application of tools from the information design literature to the study of latent representations.

Characterizing the maximum likelihood mixture estimate through the lens of sample decomposition yields key structural properties:

1. *Coherence:* The optimal sample decomposition coincides with the segmentation of the data that arises from applying Bayes' rule to the mixture estimate: it identifies which observations were likely generated by each latent type.

2. *Robustness:* Optimal latent representations exhibit local robustness. The empirical distributions within subsamples—and thus the specification of latent types—remain fixed under perturbations to the observed sample, whenever possible. Such perturbations instead lead to adjustments in the estimated prior, corresponding to changes in the relative sizes of the subsamples.

3. *Simplicity:* The optimal latent representation is simple. The number of latent types used is bounded above by the number of distinct outcomes observed in the sample or by the number of candidate data-generating processes, whichever is smaller.

We illustrate these results by showing that an estimated market segmentation remains stable under perturbations to the observed customer choices. We also demonstrate that the comparative statics of the estimated prevalence of latent types is locally simple but globally complex.

Our robustness result reinforces empirical findings that feature unobserved heterogeneity. For example, Heckman and Singer (1984) show that unemployment duration data are best explained by a heterogeneous population whose reservation wages decline over time, consistent with search theory. In contrast, a model assuming a homogeneous population would imply unrealistically increasing reservation wages. Our result implies that such substantive conclusions are robust: any alternative data formed by reweighting the latent types identified by Heckman and Singer would keep the exact same set of declining-reservation-wage latent types in play.

Section 3 introduces an additional friction, motivated by the observation that realistic economic agents may make systematic errors in their statistical analysis of observed data. We extend the model to analysts who face constraints on their ability to evaluate likelihoods. These constraints may stem from cognitive limitations (e.g., difficulty processing correlations, as in models of causal misperceptions), from institutional restrictions (e.g., fairness criteria that prohibit certain statistical classifications), or from a need to reduce the dimensionality of the problem. By limiting the analyst's ability to compute the maximum likelihood estimate, these constraints introduce a friction that is distinct from standard forms of model misspecification.

Our primary example considers an analyst who classifies customers into latent types using observed past purchases. Absent constraints, her maximum likelihood mixture estimate may reveal a correlation between the consumers' classifications and their protected characteristics, such as gender. We examine how the analyst proceeds under a statistical parity constraint requiring the classifications to be ex ante independent of gender, rendering unconstrained maximum likelihood infeasible. Our constrained sample-decomposition framework accommodates the statistical parity requirement and enables analysis of how this and related constraints influence the analyst's predictions.

The sample-decomposition constraints are technically analogous to restrictions on feasible information structures in models of persuasion or rational inattention. While such restrictions often preclude the application of standard concavification/convexification methods, we identify a property—separability—that allows these information design tools to remain applicable. We show that constraints arising from causal network structures, including those capturing limitations such as attending only to certain data dimensions or adherence to fairness criteria, satisfy this property. Consequently, our core results concerning robustness and simplicity of latent representations extend to these constrained environments. This technical insight regarding separable constraints on information structures may also contribute to the broader information design literature.

Our results also offer normative guidance for statistical practitioners who rely on ad-hoc clustering to tame dimensionality. A common workaround is to partition the data into a fixed, small number of similarity-based groups and then treat each group as if it were generated by a single latent type.[1] Yet

[1] A standard choice is the K-means algorithm, which minimizes the within-cluster sum of squared distances from the centroid. See Bonhomme et al. (2022) for its use in approximating mixture estimation, and Jehiel and Weber (2025) for its role in endogenizing the partitions in

generic clustering schemes are not robust: data perturbations can reshuffle the optimal clusters and thus change the economic meanings of the inferred types. By contrast, we derive sufficient conditions under which the constrained-optimal sample decomposition is simultaneously low-dimensional and locally robust to such perturbations.

Taken together, this paper develops a theoretical framework for understanding the formation of latent representations in the absence of exogenously specified types and priors. We reveal their inherent structure using sample decomposition and information design tools, and characterize how constraints shape these representations. Latent variables—often referred to as 'types'—are pervasive in economics but are treated inconsistently. Economic theory typically posits types as fixed primitives, unaffected by changing priors or model parameters. Conversely, statistical approaches that infer latent structure from data often yield type definitions that are inherently contingent on the specific sample observed. Our analysis bridges this gap. We show that the optimal latent representation derived from data exhibits local robustness: the inferred types remain stable under small perturbations of the data. This finding offers a microfoundation for the theoretical practice of assuming fixed types, accommodating local data variations.

2 Latent Representations

2.1 Misspecified Learning

2.1.1 Mixture Model

An analyst observes a sample $(x^i)_i$ of values from a finite set X. To abstract from sampling considerations, we focus on the limiting case of an arbitrarily large sample. We identify the sample with its empirical distribution $q_0(x)$, which is assumed to lie in the interior of $\Delta(X)$, and refer to $q_0(x)$ directly as the *sample*.[2] As in canonical learning models, the analyst assumes that the sample consists of independent and identically distributed draws from an

analogy-based equilibrium (Jehiel, 2005).

[2]Focusing on arbitrarily large samples simplifies the exposition by abstracting from sampling errors. The results in Section 2 continue to hold under finite samples. However, the optimal sample decomposition derived below typically violates integer constraints and thus must be interpreted probabilistically in finite samples. The large-sample assumption is essential in Section 3, which analyzes likelihood evaluation under frictions and relies on asymptotic arguments, as formalized in Appendix A.

underlying distribution, which she estimates via maximum likelihood.

The analyst is endowed with a compact set $\mathcal{P}_X \subseteq \Delta(X)$ of *primitive* models, where each primitive model $p(x) \in \mathcal{P}_X$ represents a data-generating process that she comprehends and deems plausible. For example, $p(x)$ may represent a stochastic choice prediction based on a particular utility specification implied by a stochastic choice theory learned in business school, or any other plausible stochastic prediction of customer choice.

The analyst, an empirical Bayesian, constructs a mixture estimate of the data-generating process using a latent variable z drawn from a set Z, perhaps identifying customers as high-elasticity, low-elasticity, or impulsive. She considers any distribution $p(z) \in \Delta(Z)$ that generates latent types z^i independently across observations i, and associates each latent type z with a conditional distribution $p(x \mid z) \in \mathcal{P}_X$ governing the distribution of x^i conditional on $z^i = z$. In this way, the analyst constructs mixture processes—distributions $p(x) = \sum_z p(z)p(x \mid z)$ in the convex hull of $\mathcal{P}_X$. We refer to a joint distribution $p(x, z) = p(z)p(x \mid z)$ as a *mixture model*.[3]

The analyst seeks a good fit to the sample as well as to infer latent characteristics of the observations. For example, a monopolist may seek to form beliefs about consumer types to target advertisements. Accordingly, she solves the *model-fitting problem*,[4]

$$\begin{aligned} \max_{\tilde{p}(x,z)} \quad & \mathrm{E}_{q_0(x)} \ln \tilde{p}(x) \\ \text{s.t.} \quad & \tilde{p}(x \mid z) \in \mathcal{P}_X \text{ for all } z \in \operatorname{supp}(\tilde{p}(z)). \end{aligned} \tag{1}$$

Importantly, the sample distribution $q_0(x)$ may fall outside $\operatorname{co}(\mathcal{P}_X)$; in that case the analyst is misspecified and cannot perfectly reproduce the data. Problem (1) is equivalent to minimization of the Kullback–Leibler divergence $\mathrm{KL}(q_0(x) \| \tilde{p}(x))$ over the same control and constraint because the two objectives differ only by a sign and a constant term—the entropy $\mathrm{H}(q_0)$ of the sample.[5]

[3]Here, the term "model" refers to the analyst's hypothesized data-generating process, in contrast to the "econometric model," which refers to the set of candidate data-generating processes, represented by $\operatorname{co}(\mathcal{P}_X)$.

[4]For existence, we assume that for each $x \in X$, there exists $p \in \mathcal{P}_X$ such that $p(x) > 0$. Then, there exists $\tilde{p} \in \operatorname{co}(\mathcal{P}_X)$ with finite $\mathrm{KL}(q_0 \| \tilde{p})$. The set of all $p' \in \operatorname{co}(\mathcal{P}_X)$ that achieve a lower value than $\tilde{p}$ does is compact. Hence, a minimizer exists due to the continuity of the objective.

[5]The Kullback–Leibler divergence $\mathrm{KL}\left(q \| p\right)$ between distributions $q(y)$ and $p(y)$, also referred to as relative entropy and commonly interpreted as a pseudo-distance, is defined as $\sum_y q(y) \ln \frac{q(y)}{p(y)}$. The Shannon entropy $\mathrm{H}(q)$ of a distribution $q(y)$ is given by $-\sum_y q(y) \ln q(y)$.

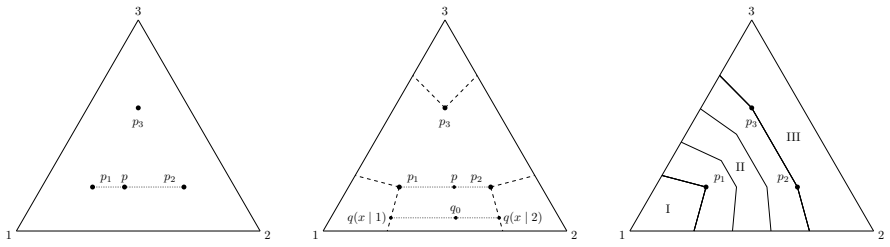


Figure 1: (A) The analyst observes choices from $X = \{1, 2, 3\}$. She constructs a mixture model p by mixing over some of the primitive models from $\mathcal{P}_X = \{p_1, p_2, p_3\}$. (B) She finds the optimal mixture model p by decomposing the sample q_0 into the subsamples $q(x \mid z = 1)$ and $q(x \mid z = 2)$ and fitting these with the primitive models p_1 and p_2, respectively. (C) $p(z) = q(z)$ depicted for $z = 1$. Region I: $p(1) = q(1) = 1$. Region III: $p(1) = q(1) = 0$. Region II: the loci depict samples with constant values of $p(1) = q(1)$.

A Bayesian approach to this problem would begin with a fixed prior $p(x, z) \in \Delta(X \times Z)$. In contrast, we are interested in analysts who lack an exogenous specification for the meaning and distribution of the latent types.

Example 1.1 (Stochastic Choice)**.** A monopolist observes a sample $q_0(x)$ of individuals' choices from the choice set X. She seeks to form beliefs about the individuals' latent types $z \in Z$, which specify individuals' utilities and possibly other choice-relevant characteristics. She subscribes to a stochastic choice theory where each latent type generates stochastic choice x by drawing from $p(x \mid z) \in \mathcal{P}_X$, as depicted in Figure 1A.

Once she has estimated the population distribution over types, $p(z)$, by fitting the sample, the monopolist applies Bayes' rule to her estimated mixture model to form posterior beliefs $p(z^i \mid x^i)$ about each individual's latent type. Combining her stochastic choice theory with these beliefs, the monopolist can generate counterfactual predictions regarding individual or aggregate choice. ▲

Throughout the paper, we adopt the standard convention $0 \ln 0 = 0$.

2.1.2 Misspecification Cost

We introduce the cost of misspecification—a concept central to our characterization of the optimal mixture model.[6] Accordingly, we refer to[7]

$$c\big(q(x);\mathcal{P}_X\big) = \min_{\tilde{p}(x)\in\mathcal{P}_X} \mathrm{KL}\left(q(x)\|\tilde{p}(x)\right)$$

as the *misspecification cost* of the sample $q(x)$ given $\mathcal{P}_X$. This measure of the mismatch between the sample and the set of considered primitive models captures the lack of fit in a hypothetical scenario where the analyst must select one of the available primitive models without mixing. When no confusion arises, we omit the argument $\mathcal{P}_X$. The following example illustrates this cost for an analyst who fails to account for correlations.

Example 2.1 (Correlation Neglect)**.** The analyst is a human resources manager who observes the education (x_1^i) and experience (x_2^i) of many employees i; thus $x = (x_1, x_2) \in X = X_1 \times X_2$. The observed sample $q_0(x_1, x_2)$ exhibits correlation. The misspecified manager fits the sample using a single primitive model $p(x_1, x_2) = p(x_1)p(x_2) \in \mathcal{P}_X = \Delta(X_1) \times \Delta(X_2)$ restricted to product distributions, without mixing. Thus, the manager is unable to capture correlations.

The maximum likelihood estimator matches the empirical marginals exactly, i.e., $p(x_k) = q_0(x_k)$, $k = 1, 2$, and induces the misspecification cost

$$c\big(q_0\big) = \mathrm{KL}\left(q_0(x_1, x_2)\|q_0(x_1)q_0(x_2)\right) = \mathrm{I}_{q_0},$$

which equals the sample's mutual information I_{q_0}. The greater the mutual information, the worse the sample is approximated by a single model assuming independence. ▲

2.2 Sample Decomposition

The model-fitting problem (1) is equivalent to decomposing the sample into subsamples so as to minimize the average misspecification cost. Working with the sample-decomposition formulation allows the analyst to employ convenient analytical tools, as we demonstrate below.

[6] See Lanzani (2025), where this cost of misspecification appears in a decision-theoretic setting from normative arguments.

[7] If $\mathcal{P}_X$ contains p such that $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$, then a minimum exists by the continuity of the objective function and the compactness of $\mathcal{P}_X$. Otherwise, the cost is infinite.

To formulate this equivalence, consider a decomposition of the sample $q_0(x)$ into subsamples, each labeled by a distinct latent variable z and having a distinct empirical distribution $q(x \mid z)$.[8] Let $q(z) \in \Delta(Z)$ denote the sample shares in the subsamples, and refer to the joint distribution $q(x, z) = q(z)q(x \mid z)$ as the *sample decomposition*. The analyst selects a sample decomposition subject to the *empirical constraint* (3), which ensures that the aggregate data corresponds to the original sample. A *sample-decomposition problem* is then defined as minimization of the overall misspecification cost:

$$\min_{\tilde{q}(x,z)} \quad \mathrm{E}_{\tilde{q}(z)}\, c\big(\tilde{q}(x \mid z); \mathcal{P}_X\big) \tag{2}$$

$$\text{s.t.} \quad \mathrm{E}_{\tilde{q}(z)}\, \tilde{q}(x \mid z) = q_0(x). \tag{3}$$

The next result shows that this formulation and the original model-fitting problem are, in fact, the same.

Proposition 1 (Decomposition Equivalence)**.** *The model-fitting problem* (1) *is equivalent to the sample-decomposition problem* (2)*. Specifically, the mixture model* $p(x, z) = p(z)p(x \mid z)$ *solves problem* (1) *if and only if*

$$p(z) = q(z), \text{ and} \tag{4}$$

$$p(x \mid z) \in \operatorname*{arg\,max}_{\tilde{p}(x) \in \mathcal{P}_X} \mathrm{E}_{q(x|z)} \ln \tilde{p}(x) \quad \forall z \in \operatorname{supp}(q(z)), \tag{5}$$

where $q(x, z) = q(z)q(x \mid z)$ *is an optimal sample decomposition that solves problem* (2)*.*

Thus, to estimate the mixture model, the analyst can first decompose the sample to minimize the average misspecification cost, solving (2), and then fit each subsample with one of the primitive models as in (5).

Establishing the equivalence in Proposition 1 requires addressing the nonlinearity of the model-fitting objective $\mathrm{KL}\big(q_0(x) \| \sum_z p(z)p(x \mid z)\big)$ from (1) with respect to $p(z)$. The proof of Proposition 1, in Appendix B.1, leverages two applications of the chain rule for the Kullback-Leibler divergence.

Both the model-fitting and sample-decomposition problems admit multiple solutions. Because the latent labels z carry no intrinsic meaning, any permutation of those labels yields another optimum. Whenever the problem exhibits

[8] If two subsamples have the same distribution, $q(x \mid z) = q(x \mid z')$, they can be pooled without loss.

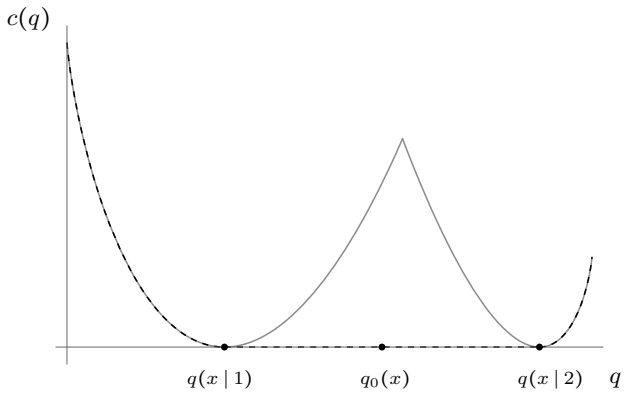


Figure 2: Sample $q_0(x)$ optimally decomposed into two subsamples. (Here, $|X| = 2$ and $\Delta(X)$ is identified with $[0, 1]$.)

nontrivial multiplicity beyond these label permutations, we let the analyst select an arbitrary solution, following the partial-identification approach. The next definition pairs a solution p of the model-fitting problem with the corresponding solution q of the sample-decomposition problem.

Definition 1. *An* optimal latent representation *is a pair comprising a mixture model $p(x, z)$ and a sample decomposition $q(x, z)$, where q solves the sample-decomposition problem* (2) *and the corresponding optimal mixture model p satisfies* (4) *and* (5)*.*

The sample-decomposition problem (2) is analogous to the concavification problem of Kamenica and Gentzkow (2011), except that we minimize rather than maximize the objective, and hence 'convexify' rather than concavify the objective function. Accordingly, let $C(q)$ denote the convex envelope of the cost function $c(q)$.[9] For a sample $q_0(x)$, the *tangency points of the convexification* are the distributions q such that $(q, c(q))$ lies on the supporting hyperplane to the graph of the convex envelope C at $(q_0, C(q_0))$. The weights used in expressing q_0 as a convex combination of these tangency points constitute the *distribution over the tangency points*. The sample decomposition $q(x, z)$ solves problem (2) if and only if the subsamples $q(x \mid z)$ are tangency points of the convexification and $q(z)$ is the corresponding distribution over these tangency points. Figure 2 illustrates.

[9] Recall that the convex envelope of a function $c(q)$ is a function $C(q) = \inf\{\xi : (q, \xi) \in \mathrm{co}(c)\}$, where $\mathrm{co}(c)$ stands for the convex hull of the graph of c.

Example 1.2 (Sample Decomposition for Stochastic Choice)**.** We return to the monopolist's problem from Figure 1A. When the sample q_0 lies within the convex hull of the primitive model set $\mathcal{P}_X$, the monopolist achieves the perfect fit $p(x) = q_0(x)$ by decomposing the sample into three subsamples $q(x \mid z) = p_z(x)$, $z = 1, 2, 3$, each of which coincides with one of the primitive models. In this simple case, a local perturbation of the sample q_0 alters only the distribution $p(z) = q(z)$ while leaving the optimal subsamples $q(x \mid z)$ intact.

When the sample q_0 lies outside the convex hull of $\mathcal{P}_X$, the monopolist cannot perfectly match the sample, resulting in misspecification. For the sample q_0 depicted in Figure 1B, the analyst optimally decomposes the sample into two subsamples, $q(x \mid z = 1)$ and $q(x \mid z = 2)$, and fits each subsample with the closest primitive model, p_1 or p_2. In this case, the effect of a local perturbation of the observed sample depends on the perturbation. If $q_0(x)$ remains within the convex hull of the two original subsamples, the optimal subsamples remain unchanged, and only the distribution of the latent type adjusts. Conversely, changes in $q_0(x)$ that extend outside this convex hull alter the subsamples' distributions.

Finally, for completeness, if the sample q_0 assigns nearly all probability to a single value of x, corresponding to samples in the neighborhoods of the simplex vertices (as delineated by the dashed lines in Figure 1B), the analyst concludes that all sampled individuals belong to the nearest type p_z, effectively collapsing the latent structure to a single type.

We solved this example by exploiting a mathematical equivalence between the sample-decomposition problem and the rational inattention problem introduced by Matějka and McKay (2015). By associating each primitive distribution $p(x) \in \mathcal{P}_X$ with a unique label $a \in A$ and defining a utility function $u(a, x) = \ln p_a(x)$, we can reformulate the sample-decomposition problem (2) as a standard rational-inattention problem of the form:

$$\begin{aligned} \max_{\tilde{q}(x,z)} \quad & \mathrm{E}_{\tilde{q}(z)} \left[\max_{a \in A} \mathrm{E}_{\tilde{q}(x|z)} \, u(a, x) + \mathrm{H}\left(\tilde{q}(x \mid z)\right) \right] \\ \text{s.t.} \quad & \mathrm{E}_{\tilde{q}(z)} \, \tilde{q}(x \mid z) = q_0(x). \end{aligned}$$

Hence, the solutions to the sample-decomposition problems correspond to the solutions to the rational-inattention problems with utility $u(a, x) = \ln p_a(x)$. Using this analogy, we adapted the solution for the rational inattention problem with three states and actions from Matyskova and Montes (2023). ▲

2.3 Optimal Latent Representations

We first show that optimal latent representations are internally coherent, robust to sample perturbations, and simple. Finally, Subsection 2.3.4 explains how the analyst updates her belief about the prevalence of latent types when the sample changes.

2.3.1 Coherence

The mixture model $p(x, z)$ represents the analyst's estimate of the data-generating process. In contrast, the sample decomposition $q(x, z)$, which emerges as a side product of estimation, represents the analyst's organization of the data. If the analyst is well-specified, then the mixture model p and sample decomposition q coincide. When the analyst is misspecified, $p(x, z)$ and $q(x, z)$ differ because the analyst fails to perfectly fit the sample distribution; $p(x) \neq q(x) \equiv q_0(x)$. Nonetheless, the analyst's model and her sample decomposition provide mutually coherent accounts of the latent variables. The following result was established in Proposition 1 and its proof.

Corollary 1. *The analyst assigns latent values to observations coherently with her mixture model:* $q(z) = p(z)$ *and* $q(z \mid x) = p(z \mid x)$.

To illustrate, consider the monopolist from Example 1 who infers individuals' latent types from their observed choices. This analyst may adopt one of two conceptually distinct inference procedures. The first procedure involves explicit Bayesian reasoning, as in the empirical Bayes approach of Robbins (1956). The analyst first estimates the joint population distribution $p(x, z)$ over latent types and choices and then computes the Bayesian beliefs $p(z \mid x)$ for each individual given their observed choice. The second procedure does not involve explicit Bayesian updating. In this approach, the analyst decomposes the sample to minimize the misspecification cost and then samples individuals who made choice x to learn the empirical distribution $q(z \mid x)$ of the labels z they were attributed within the sample decomposition. The corollary states that the two procedures generate the same beliefs. An analyst who has found it convenient to address her estimation problem by solving the sample decomposition problem can evaluate latent types by working directly with the sample decomposition $q(x, z)$.

Similarly, an analyst interested in the counterfactual prediction of aggregate behavior may form beliefs about the population distribution of the latent types using two distinct procedures. She can marginalize her estimated mixture

model $p(x, z)$ to obtain the belief $p(z)$. Alternatively, she can draw from her sample decomposition to learn the empirical distribution $q(z)$. Again, these two procedures yield identical beliefs.

Additionally, Corollary 1 implies that the optimal subsamples $q(x \mid z)$ admit a Bayesian reinterpretation. The analyst's within-sample Bayesian prediction of the observable outcome of each latent type z equals the subsample distribution $q(x \mid z)$. To see this, recall that an analyst who observes the empirical distribution $q_0(x)$ and selects the mixture model $p(x, z)$ assigns joint beliefs over (x, z) within her sample as

$$q_0(x)\, p(z \mid x) \;=\; q_0(x)\, q(z \mid x) \;=\; q(x, z).$$

Conditioning on the latent type z yields the result.

2.3.2 Robustness

How does the analyst adapt her optimal latent representation in response to changes in the observed sample? In principle, the analyst may respond by modifying the subsample distributions, adjusting the primitive models fitted to those subsamples, or updating the distribution over latent types. We find that, whenever feasible, the analyst prioritizes adjusting the distribution over latent types, while keeping both the subsample distributions and their fits intact.

Definition 2 (Local Robustness)**.** *We say that an optimal latent representation exhibits* local robustness *if, whenever the sample changes within the convex hull of the optimal subsamples, the subsamples remain optimal. Formally, for any optimal latent representation $p(x, z)$ and $q(x, z)$ associated with the sample $q_0(x)$ and any distribution $r(z)$ with $\operatorname{supp}(r(z)) \subseteq \operatorname{supp}(p(z))$, the latent representation given by $p'(x, z) = r(z)p(x \mid z)$ and $q'(x, z) = r(z)q(x \mid z)$ is optimal for the perturbed sample $q_0'(x) = \sum_z r(z)q(x \mid z)$.*

Since the tangency points of convexification are invariant to local perturbations of the sample, it follows from Proposition 1 that optimal latent representations are locally robust. Appendix B.2 proves:

Corollary 2. *Optimal latent representations exhibit local robustness.*

This robustness implies, for example, that the market segmentation adopted by a firm need not be reconfigured in response to local variations in the firm's market research data. Recall the monopolist who observes choices x^i made by

customers i in her market. She estimates a mixture model $p(x, z)$ and then forms within-sample Bayesian predictions about the choice of her customer, conditional on that customer's type. Since her within-sample prediction for type z's choice is $q(x \mid z)$—the monopolist decomposes the aggregate demand into type-specific segments $q(x \mid z)$, with corresponding market shares $q(z)$. Corollary 2 states that this market segmentation continues to be applicable in the face of local variations in market conditions. That is, if the observed sample changes within the convex hull of the original type-specific segments, then each segment $q(x \mid z)$ remains unmodified, and the monopolist adjusts only the market shares $q(z)$.

2.3.3 Simplicity

Optimal latent representations satisfy two simplicity properties. First, Carathéodory's theorem implies a bound on the number of employed latent types.[10]

Corollary 3. *There exists an optimal latent representation that employs at most as many latent types as the number $|X|$ of observed outcomes: $|\text{supp}(q(z))| = |\text{supp}(p(z))| \leq |X|$.*

Lindsay (1983) derives the same bound using a related but distinct convex geometry argument. Our Corollary 7 below provides sharper bounds in settings with estimation frictions.[11]

Second, in any optimal latent representation, each latent type has a distinct meaning, as formalized in the next definition. This distinctness implies a further upper bound on the number of employed latent types.

Definition 3. *A latent representation is* parsimonious *if $p(x \mid z) \neq p(x \mid z')$ for all distinct types $z, z' \in \text{supp}(q(z))$.*

A failure of parsimony arises when two latent types z and z' have identical predictive content—i.e., $p(x \mid z) = p(x \mid z')$—yet the analyst assigns them distinct

[10] We apply the same argument used to bound the support of the optimal signal in the Bayesian persuasion and rational inattention literatures. Since q_0 lies in the convex hull of at most $(|X| - 1)$-dimensional set $\{q(x \mid z)\}_z$ of optimal subsamples for some optimal decomposition q, it can be expressed as a convex combination of at most $|X|$ such elements, and this reduced convex combination achieves the same objective value as the original sample decomposition.

[11] Lindsay represents each primitive model $p(x) \in \mathcal{P}_X$ as an $(|X|-1)$-dimensional vector specifying the likelihood associated with each $x \in X$, with each mixture process lying in the convex hull of $\mathcal{P}_X$. Lindsay derives his bound by applying Carathéodory's theorem to this convex hull. In contrast, we apply Carathéodory's theorem to the optimal sample decompositions. Our approach, relative to Lindsay's, provides insights into the latent-type decomposition and tightens the bound on the employed latent types when the decomposition is constrained.

subsamples $q(x \mid z) \neq q(x \mid z')$. Parsimony implies that the analyst employs at most as many latent types as there are available primitive models, yielding a nontrivial bound when $|\mathcal{P}_X|$ is finite.

Corollary 4. *Any optimal latent representation is parsimonious. Hence, it employs at most $|\mathcal{P}_X|$ latent types: $|\operatorname{supp}(q(z))| = |\operatorname{supp}(p(z))| \leq |\mathcal{P}_X|$.*

A failure of parsimony cannot arise in the current setting, since the analyst would strictly benefit from pooling the two subsamples, as shown in the proof of the corollary in Appendix B.3. However, parsimony may fail under additional constraints introduced in Section 3, as illustrated in Example 4.

2.3.4 Updating the Latent-Type Prior

The empirical Bayes approach predicts how an analyst updates her prior $p(z) = q(z)$—chosen to fit the observed sample $q_0(x)$—when $q_0(x)$ is perturbed. This updating task is distinctly non-Bayesian: rather than applying Bayes' law to a fixed prior, the analyst adjusts the prior to fit the perturbed data.[12]

Corollary 2 implies that the empirical prior is locally linear in the observed sample:

Corollary 5. *The empirical prior $p(z) = q(z)$ is locally linear in the sample $q_0(x)$: Suppose $q(x, z)$ is an optimal sample decomposition of $q_0(x)$. Let $p'(z) = q'(z)$ be any distribution with $\operatorname{supp}(p'(z)) \subseteq \operatorname{supp}(q(z))$. Then $p'(z) = q'(z)$ is the empirical-Bayes prior for the perturbed sample $q'_0(x) = \mathrm{E}_{p'(z)}\, q(x \mid z)$.*

While the relationship between the empirical prior and the sample is locally linear, it gives rise to a rich comparative statics globally, manifesting as over- and under-reactions relative to a benchmark. We compare the empirical Bayes analyst to a benchmark, which we refer to as an *ideational* Bayesian. This benchmark analyst holds a fixed prior $p(x, z)$ and computes the within-sample prevalence of type z as

$$q_{\mathrm{ib}}(z) = \mathrm{E}_{q_0(x)}\, p(z \mid x), \tag{6}$$

which is *globally* linear in $q_0(x)$. Let us contrast $q_{\mathrm{ib}}(z)$ with the empirical Bayes estimate $q(z) = p(z)$.

[12] Bervoets et al. (2025) link the empirical Bayes updating task studied here to standard Bayesian updating. In a sequential-sampling setting, they show that an updating rule formed by a convex combination of the Bayesian update and the current prior—with a weight on the Bayesian update that declines over time—converges to the empirical-Bayes prior $p(z)$.

Compared to $q_{\text{ib}}(z)$, the empirical Bayes estimate $q(z)$ can exhibit both overreaction and underreaction to variations in the observed sample $q_0(x)$. To illustrate overreaction, consider an increase in the prevalence of an observation x^* (e.g., selling an asset) that is stochastically indicative of a particular latent type z^* (e.g., a pessimistic investor). Under both the ideational and empirical Bayes approaches, this shift in $q_0(x)$ leads to a direct increase in the estimated prevalence of pessimistic investors, as captured in (6). Additionally, under the empirical Bayes approach, the posteriors $p(z \mid x) = q(z \mid x)$ shift toward the pessimistic type z^* for all observations, amplifying the response and inducing a relative overreaction. In contrast, underreaction arises for extreme samples—such as those in the corner regions of Figure 1B—when the empirical Bayes estimate becomes locally insensitive to $q_0(x)$. See Figure 1C for illustration.

Because the mapping from the sample $q_0(x)$ into $q(z)$ is not globally linear, the empirical Bayes estimate $q(z)$ can fail an analogue of the law of iterated expectations. First, suppose the analyst observes samples $q_0^A(x)$ and $q_0^B(x)$ from two distinct subpopulations comprising proportions λ and $1-\lambda$ of the overall population. She estimates the latent type prevalences in each subpopulation, and forms the corresponding aggregate estimate $\lambda q^A(z) + (1-\lambda)q^B(z)$. Now suppose instead that the analyst observes only the aggregate sample $q_0^M(x) = \lambda q_0^A(x) + (1-\lambda)q_0^B(x)$ and forms directly the corresponding estimate $q^M(z)$. The non-linearity implies $q^M(z) \neq \lambda q^A(z) + (1-\lambda)q^B(z)$, opening the door to manipulation via strategic partitioning of the sample.

3 Constrained Likelihood Evaluation

We have hitherto assumed that the analyst is constrained to mixtures of primitive models from $\mathcal{P}_X$—and may therefore be misspecified—but can nonetheless evaluate the fit of mixture models without friction. The following example illustrates that, in some settings, when no frictions are imposed on the likelihood evaluation, the analyst may, somewhat unrealistically, circumvent misspecification frictions altogether by leveraging a sufficiently high-dimensional latent representation.

Example 2.2 (Correlation Despite Correlation Neglect)**.** Recall the human resources manager from Example 2.1, who is constrained to primitive models $p(x_1, x_2) = p(x_1)p(x_2)$ that satisfy independence. By mixing over such models, the manager can perfectly fit any sample $q_0(x_1, x_2)$ regardless of its correlation.

This is achieved by decomposing $q_0(x_1, x_2)$ into degenerate subsamples, each of which places all probability on a single pair (x_1, x_2). Each such subsample is then fitted with a degenerate primitive model that trivially satisfies the independence restriction. ▲

In this example, the manager achieves a perfect fit—despite limitations on the set of primitive models, which may be intended to reflect complexity constraints—by employing and flawlessly evaluating a sufficiently high-dimensional mixture model. For constraints such as correlation neglect to be consequential, some additional friction, such as dimensionality restrictions, must arise. This section introduces internal and external frictions in likelihood evaluation and derives the resulting constrained-optimal latent representations.

3.1 Constrained Sample Decomposition

To introduce friction into the likelihood evaluation of mixture models, we introduce a compact set

$$\mathcal{Q} \subseteq \big\{q(x, z) : q(x) = q_0(x)\big\} \tag{7}$$

of admissible sample decompositions, which we refer to as the *decomposition constraint*. We generalize the sample-decomposition problem (2) by introducing the *constrained* sample-decomposition problem:

$$\min_{\tilde{q}(x,z)\in\mathcal{Q}} \mathrm{E}_{\tilde{q}(z)}\, c\big(\tilde{q}(x \mid z); \mathcal{P}_X\big). \tag{8}$$

Since $\mathcal{Q}$ subsumes the empirical constraint, this indeed generalizes problem (2).

Decomposition constraints may arise from internal cognitive limitations or from external institutional constraints. The following example illustrates the latter case.

Example 1.3 (Market Segmentation meets Statistical Parity)**.** Recall the monopolist who observes choices of individuals i, denoted here by $x_1^i \in X_1$ to distinguish them from other observed characteristics $x_0^i \in X_0$. The monopolist's stochastic choice theory is summarized by the set $\mathcal{P}_X \subseteq \Delta(X_1)$ of primitive models that involve only the variable x_1. She assigns each individual i a latent type z^i via a stochastic classification rule $q(z \mid x_1)$. This rule, together with the sample $q_0(x_1)$, implies a sample decomposition $q(x_1, z) = q_0(x_1)q(z \mid x_1)$. As before, the monopolist minimizes the average misspecification cost $\mathrm{E}_{q(z)}\, c\big(q(x_1 \mid z); \mathcal{P}_X\big)$.

Suppose now that the classification is constrained by another observed characteristic x_0^i, which acts as a *spoiler variable*. Specifically, assume the spoiler variable x_0^i represents a protected demographic characteristic—taken here to be gender. If unconstrained, the monopolist would ignore the spoiler variable x_0, as it plays no role in her stochastic choice theory. However, the monopolist is externally constrained to comply with statistical parity fairness, which requires the classification be a priori independent of the protected attribute.[13]

Let $q_0(x_0, x_1)$ be the joint sample distribution over gender and choice, and let $q(x_0, x_1, z)$ be the joint distribution of gender, choice, and the classifier's output. The statistical parity constraint restricts the monopolist to choose $q(x_0, x_1, z)$ from

$$\tilde{\mathcal{Q}} = \big\{q(x_0, x_1, z) : z \perp\!\!\!\perp_q x_0 \text{ and } q(x_0, x_1) = q_0(x_0, x_1)\big\},$$

inducing the decomposition constraint $\mathcal{Q} = \big\{q(x_1, z) : q(x_0, x_1, z) \in \tilde{\mathcal{Q}}\big\}$. ▲

Appendix A presents a microfoundation of the constrained sample-decomposition problem. In short, we define the *constrained divergence* of a model $\tilde{p}(x, z)$ from the observed sample $q_0(x)$,[14]

$$\mathrm{D}_{\mathcal{Q}}\big(\tilde{p}(x, z)\big) = \min_{\tilde{q}(x,z)\in\mathcal{Q}} \mathrm{KL}\big(\tilde{q}(x, z) \parallel \tilde{p}(x, z)\big), \tag{9}$$

and note that it coincides with the standard measure of misfit, $\mathrm{KL}\big(\tilde{q}(x) \parallel \tilde{p}(x)\big)$, if the decomposition constraint $\mathcal{Q}$ is relaxed to the empirical constraint $q(x) = q_0(x)$. We show that the constrained divergence is equivalent to the constrained likelihood of the mixture model $\tilde{p}(x, z)$. We then introduce the constrained model-fitting problem, in which the analyst optimizes over feasible mixture models $\tilde{p}(x, z)$ to minimize the constrained divergence $\mathrm{D}_{\mathcal{Q}}$, and prove that this problem is equivalent to the constrained sample-decomposition problem (8). Analogously to the unconstrained case, we refer to the optimized pair (p, q) as a constrained-optimal latent representation.

The decomposition constraint is conceptually distinct from misspecification. Misspecification restricts the set of models the analyst considers. The decom-

[13]Statistical parity fairness, a leading fairness concept from the field of machine learning, requires that the classifier's output be a priori independent of a protected demographic characteristic (Corbett-Davies et al., 2017). A breach of statistical parity may indicate potential disparate impact, which could trigger legal scrutiny. See also Strack and Yang (2024) and He et al. (2021), who study information structures that are constrained to not provide information about either certain aspects of the underlying state or about other signals.

[14]If there exists a $q(x, z) \in \mathcal{Q}$ such that $\mathrm{supp}(q) \subseteq \mathrm{supp}(\tilde{p})$, then a finite minimizer exists. Otherwise, we define $\mathrm{D}_{\mathcal{Q}}$ to be infinite.

position captures limitations on her ability to evaluate the likelihoods of those models.

3.2 Constrained-Optimal Latent Representations

We now examine to what extent, and under what conditions, the properties of internal coherence, robustness, and simplicity carry over from the unconstrained setting to constrained-optimal latent representations.

3.2.1 (Approximate) Coherence

The internal coherence of optimal latent representations extends to the constrained setting, albeit in a partially approximate form.

Corollary 6. *The analyst assigns latent values to observations in an internally coherent manner at the aggregate level:*

$$q(z) = p(z).$$

At the disaggregated level, the analyst's stochastic assignment of latent values, $\big(q(z \mid x)\big)_x$*, approximates the Bayesian updates* $\big(p(z \mid x)\big)_x$ *of the mixture model by solving the following projection problem:*

$$\big(q(z \mid x)\big)_x \in \underset{\left(\tilde{q}(z|x)\right)_x}{\arg\min} \quad \mathrm{E}_{q_0(x)} \, \mathrm{KL}\left(\tilde{q}(z \mid x) \| p(z \mid x)\right)$$
$$\text{s.t.} \quad q_0(x)\tilde{q}(z \mid x) \in \mathcal{Q}.$$

That is, although in the constrained setting, the analyst's classification rule $q(z \mid x)$ is generally incoherent with her own Bayesian belief $p(z \mid x)$, this incoherence is minimized subject to the decomposition constraint.

Constrained-optimal latent representations yield internally coherent aggregate predictions, offering a novel foundation for the rational expectations hypothesis. To see the connection, rewrite $p(z) = q(z)$ as

$$p(z) = \mathrm{E}_{q_0(x)} \, q(z \mid x), \tag{10}$$

and consider the analyst who classifies individuals using the stochastic classification rule $q(z \mid x)$ that conditions on the observed choices x. Equation (10)

states that the analyst's endogenously chosen prior $p(z)$ is coherent with the actual classification probabilities $\mathrm{E}_{q_0(x)}\, q(z \mid x)$: the analyst is not systematically surprised when testing her prior against her own classification of the sample.

The condition in (10) differs from the standard notion of Bayesian plausibility, which similarly requires that the average updated belief equals the prior belief. Bayesian plausibility is an identity within a single joint distribution, enforced by the mechanics of Bayesian updating and independent of any optimality considerations. In contrast, the rational-expectations condition (10) relates two distinct distributions, is not an identity, and can fail. Nonetheless, it holds for constrained-optimal latent representations.

A popular intuition supporting rational expectations suggests that an analyst who is systematically surprised should revise her prior to eliminate the discrepancy. While this intuition does not fit within the standard Bayesian framework, where the prior is fixed, it aligns naturally with our approach. Indeed, our analyst maximizes the constrained likelihood by adjusting $p(z)$ to match the empirical average of the classification probabilities $q(z \mid x)$.[15]

3.2.2 Separable Decomposition Constraints

We aim to extend the robustness and simplicity properties of unconstrained-optimal latent representations to the constrained setting. Since these properties are implied by convexification, we restrict attention to a class of decomposition constraints that preserve convexification as a valid solution method. We illustrate the failure of robustness in Example 3 below, then introduce our class of decomposition constraints that preserves robustness in an abstract form. We show in Section 3.2.3 that this class includes decomposition constraints of applied relevance.

Example 3 (Fragility of Variational Bayes Methods)**.** Practitioners often resort to Variational Bayes methods when exact Bayesian updating or maximum-likelihood estimation is numerically infeasible; these methods are a prominent special case of our constrained sample-decomposition problem. These methods, as developed by Jordan et al. (1999) and Kingma and Welling (2013), restrict the conditional distributions $q(z \mid x)$ to lie in a set $\mathcal{Q}_Z \subset \Delta(Z)$, thereby inducing

[15] Spiegler (2020) provides sufficient conditions for rational expectations in a related, though non-nested, bounded-rationality framework. Like us, Spiegler defines rational expectations as a match between the true average of the analyst's subjective posterior beliefs and her subjective prior. While Spiegler relies on a structure of directed acyclic graphs, we do not assume this structure for this result.

the decomposition constraint[16]

$$\mathcal{Q} = \big\{q(x,z) : q(z \mid x) \in \mathcal{Q}_Z \text{ and } q(x) = q_0(x)\big\}.$$

Each $q(z \mid x)$ is interpreted as an approximation to the Bayesian update $p(z \mid x)$, and the constrained–optimal mixture model $p(x,z)$ then serves as an approximation to the exact maximum-likelihood mixture estimate.

Nevertheless, the Variational Bayes approach is *not* robust: a perturbation of the sample $q_0(x)$ typically alters both the subsamples $q(x \mid z)$ and the inferred latent types $p(x \mid z)$. We address this by introducing *separable* decomposition constraints that ensure robustness under sample perturbations. ▲

To introduce separable decomposition constraints, consider an analyst who observes a sample $q_0(x_0, \dots, x_K)$, where each observation x^i is a tuple $(x_0^i, \dots, x_K^i)$. The analyst seeks to learn the joint distribution of $(x_1, \dots, x_K)$. As in Example 1.3, the additional variable x_0 serves as a *spoiler variable*, potentially influencing the sample decomposition. Since the analyst aims to fit the distribution of only the variables $x_1, \dots, x_K$, the constrained sample-decomposition problem becomes

$$\min_{\tilde{q}(x_0,\dots,x_K,z)\in\mathcal{Q}} \mathrm{E}_{\tilde{q}(z)}\, c\big(\tilde{q}(x_1,\dots,x_K \mid z); \mathcal{P}_X\big), \tag{11}$$

where the misspecification cost depends solely on the distribution of the variables $x_1, \dots, x_K$ that the analyst seeks to estimate, while the spoiler variable x_0 enters the problem only through the decomposition constraint.[17]

Definition 4. *A decomposition constraint $\mathcal{Q} \subseteq \Delta(X \times Z)$, with $X = X_0 \times X_1 \times \dots X_K$, is said to be* separable with marginal constraint $\mathcal{Q}_X \subseteq \Delta(X)$ *if*

$$\mathcal{Q} = \big\{q(x,z) : q(x \mid z) \in \mathcal{Q}_X \text{ for all } z \in \operatorname{supp}(q(z)) \text{ and } q(x) = q_0(x)\big\}.$$

A decomposition constraint is separable if it restricts each subsample to a set $\mathcal{Q}_X$, while imposing no additional global constraints beyond the empirical constraint.

Separable constraints permit a straightforward extension of the convexification method. Instead of convexifying the misspecification cost function, the

[16] E.g., the popular *mean-field* specification takes a multidimensional latent variable $z = (z_1, \dots, z_K)$ and assumes conditional independence, so that $\mathcal{Q}_Z = \Delta(Z_1) \times \dots \times \Delta(Z_K)$.

[17] We adopt a slight abuse of notation by writing $\mathcal{Q}$ for the constraint over joint distributions $q(x_0, \dots, x_K, z)$. More precisely, we should distinguish between the constraint $\tilde{\mathcal{Q}}$ over $q(x_0, \dots, x_K, z)$ and the marginalized constraint $\mathcal{Q}$ over $q(x_1, \dots, x_K, z)$. However, since we work exclusively with the first constraint, we economize notation by denoting it simply as $\mathcal{Q}$.

extended approach convexifies an augmented cost function that penalizes infeasible subsamples with an infinite penalty. Thus, the constrained sample-decomposition problem (11) is equivalent to the convexification problem

$$\begin{aligned} \min_{\tilde{q}(x_0,\ldots,x_K,z)} \quad & \mathrm{E}_{\tilde{q}(z)}\, \tilde{c}\big(\tilde{q}(x_0,\ldots,x_K \mid z)\big) \\ \text{s.t.} \quad & \mathrm{E}_{\tilde{q}(z)}\, \tilde{q}(x_0,\ldots,x_K \mid z) = q_0(x_0,\ldots,x_K), \end{aligned}$$

with the augmented cost function

$$\tilde{c}\big(q(x_0,\ldots,x_K)\big) = \begin{cases} c\big(q(x_1,\ldots,x_K);\mathcal{P}_X\big) & \text{if } q(x) \in \mathcal{Q}_X, \\ \infty & \text{otherwise.} \end{cases}$$

3.2.3 DAG-Based Constraints

We now introduce a natural class of separable decomposition constraints.

Many decomposition constraints can be represented using directed acyclic graphs (DAGs), which encode conditional independence assumptions through a graph structure. For instance, recall the monopolist from Example 1.3, who assesses individuals' latent types from their observed choices (x_1), while accounting for gender (x_0) as a spoiler variable. The statistical parity constraint—which requires that classification be independent of gender—is equivalent to requiring that $q(x_0, x_1, z)$ factorize according to the DAG

$$z \to x_1 \leftarrow x_0, \tag{12}$$

with the induced factorization constraint

$$q(x_0, x_1, z) = q(z)q(x_0)q(x_1 \mid x_0, z).$$

More generally, we use DAGs to represent the analyst's capacity to correlate the latent variable z with the observables $(x_0,\ldots,x_K)$. We restrict the analyst to sample decompositions $q(x,z)$ that factorize as

$$q(x,z) = q(z)\prod_{k=0}^{K} q\big(x_k \mid \mathrm{pa}(x_k)\big), \tag{13}$$

where $\mathrm{pa}(x_k)$ is the set of parents of the variable x_k in the given DAG. That is, the DAG encodes a system of conditional independence assumptions over

the random variables x and z. Specifically, it requires that each variable be conditionally independent of its non-descendants given its immediate parents in the graph.[18]

We allow for arbitrary DAG in which the latent variable z is a root node, meaning it has no parents. This assumption means that the analyst views the latent variable as one of the primary causes of the observables.[19]

We say that the decomposition constraint $\mathcal{Q}$ is *DAG-based* if $\mathcal{Q}$ consists of all sample decompositions $q(x, z)$ that (i) factorize according to (13) and (ii) satisfy the empirical constraint $q(x) = q_0(x)$. We assume that the DAG-based constraint $\mathcal{Q}$ is non-empty; that is, the sample $q_0(x)$ is compatible with the DAG. Appendix B.4 proves:

Lemma 1. *The DAG-based constraint $\mathcal{Q}$ is separable, with marginal constraint $\mathcal{Q}_X$ equal to the set of distributions $q(x)$ that:*

(i) factorize as $q(x) = \prod_{k=0}^{K} q\big(x_k \mid \mathrm{pa}'(x_k)\big)$, where $\mathrm{pa}'(x_k) = \mathrm{pa}(x_k) \setminus \{z\}$, and

(ii) $q\big(x_k \mid \mathrm{pa}(x_k)\big) = q_0\big(x_k \mid \mathrm{pa}(x_k)\big)$ for all k such that $z \notin \mathrm{pa}(x_k)$.

That is, the feasible subsamples $q(x \mid z) \in \mathcal{Q}_X$ (i) inherit the factorization from the DAG, and (ii) have the conditional distributions $q\big(x_k \mid \mathrm{pa}(x_k), z\big)$ anchored to the empirical conditionals $q_0\big(x_k \mid \mathrm{pa}(x_k)\big)$ for all variables x_k that the DAG specifies as independent of z.

The next example illustrates the separability of the DAG-based constraints and its implications.

Example 1.4 (Constrained-Optimal Market Segmentation)**.** Recall the monopolist who classifies individuals based on their observed choices (x_1) and is constrained to classifications that are independent of gender (x_0). This statistical parity constraint is represented by the DAG-based constraint (12). We now characterize the resulting prediction errors.

First, observe that the DAG in (12) alone imposes a *global* constraint, $q(x_0 \mid z) = q(x_0 \mid z')$ for any z and z', and recall that global constraints preclude the

[18] For early economic applications of DAGs to bounded rationality, see Spiegler (2016). The classical reference is Pearl (1988).

[19] The assumption that z is a root node is crucial for the separability of the DAG-based constraints, since such DAGs do not impose restrictions on the marginal distribution $q(z)$. In contrast, general DAGs may impose restrictions on $q(z)$ given $q(x \mid z)$, thereby violating separability.

convexification method. However, the intersection of this constraint with the empirical constraint is *separable* since together they imply $q(x_0 \mid z) = q_0(x_0)$ for each z. Thus, a sample decomposition $q(x, z)$ is feasible if and only if $q(x_0, x_1 \mid z) \in \mathcal{Q}_X$ for each z with the marginal constraint $\mathcal{Q}_X = \{q(x_0, x_1) : q(x_0) = q_0(x_0)\}$.

For simplicity, assume that individuals make binary choices, where $x_1 = 1$ indicates a purchase and $x_1 = 0$ indicates abstention. We refer to various (conditional) purchase probabilities as demands. Let the sample have equal proportions of men ($x_0 = m$) and women ($x_0 = w$). The monopolist assumes that each individual is either a high ($z = h$) or low ($z = l$) latent type, with purchase occurring with probabilities $p_h > p_l = 1 - p_h$, corresponding to the set of primitive models $\mathcal{P}_X = \{p_h(x_1), p_l(x_1)\}$. Under this specification, the constrained model-fitting problem reduces to estimating the share of high types in the sample.

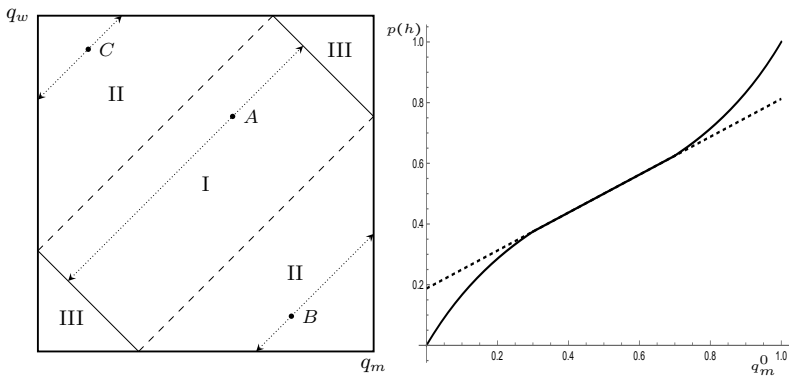
For a benchmark, we temporarily lift the statistical parity constraint, allowing the monopolist to ignore gender. The monopolist then achieves a perfect fit whenever the observed demand $q_0(x_1)$ is between p_l and p_h by decomposing the sample into two subsamples with their demands matching the demands of the two types (otherwise, the monopolist assumes that all individuals are of the same best-fitting type).

We now reintroduce the statistical parity constraint, forcing the monopolist to engage with gender. Since the gender distributions of the subsamples are fixed at $q(x_0 \mid z) = q_0(x_0)$, we can identify each feasible subsample $q(x_0, x_1 \mid z) \in \mathcal{Q}_X$ with a pair $(q_m^z, q_w^z) \in [0,1]^2$ that specifies the gender-specific demands $q_m^z := q(x_1 = 1 \mid x_0 = m, z)$ and $q_w^z := q(x_1 = 1 \mid x_0 = w, z)$ within the subsample. Similarly, the monopolist represents the observed sample $q_0(x_0, x_1)$ by its gender-dependent demands (q_m^0, q_w^0). Using this representation, the constrained sample-decomposition problem becomes equivalent to unconstrained decomposition of (q_m^0, q_w^0) into a convex combination of $\left(q_m^z, q_w^z\right)_z$ that minimizes the expectation of the cost

$$c\big(q_m, q_w\big) = \min_{z \in \{l,h\}} \mathrm{KL}\left(\frac{q_m + q_w}{2} \| p_z\right).$$

Finally, the monopolist assigns $p(x_1 \mid z) = p_h(x_1)$ to any subsample with average demand $\frac{q_m^z + q_w^z}{2} > 1/2$ and assigns $p_l(x_1)$ otherwise. See Figure 3A for the constrained-optimal sample decomposition.

Figure 3B illustrates the monopolist's constrained prediction of the high-type



(A) (B)

Figure 3: (A) Constrained-optimal sample decomposition: For samples in region I with moderate gender differences in demands, the statistical parity constraint is slack, and the subsamples' average demands $(q_m^z + q_w^z)/2$ align with either the low or high type's demand, p_l or p_h (e.g., sample A). In region II, the constraint binds, as illustrated by samples B and C. Finally, in region III, the monopolist fits the sample with a single type.
(B) Full curve: Predicted share $p(z) = q(z)$ of high-type individuals ($z = h$) as a function of men's demand q_m^0. Women's demand is fixed at $q_w^0 = 0.5$. Dashed line: the benchmark unconstrained prediction of the high type's share.
(The demands of the high and low latent types are parameterized as $p_h = 0.9$ and $p_l = 0.1$.)

prevalence, $p(z)$, $z = h$, as a function of the gender-specific demands observed in the sample. When the gender differences in the demands are moderate, the monopolist's constrained prediction coincides with the benchmark unconstrained prediction. However, for strong enough gender differences, the monopolist's constrained prediction deviates from the benchmark, resulting in a poorer fit.

For intuition, observe that the monopolist aims to partition individuals into two subsamples, each corresponding to one of the two assumed latent types characterized by high and low demand. The monopolist faces the constraint that this partition must not indirectly reveal gender information. When gender differences in demand are mild, the constraint is slack. As gender differences increase, partitioning individuals based on purchases becomes progressively restricted, reducing model fit. ▲

3.2.4 Robustness

The robustness of optimal latent representations extends to settings with DAG-based decomposition constraints.

Proposition 2. *If $\mathcal{Q}$ is a DAG-based constraint, then optimal latent representations exhibit local robustness.*

The proof of the proposition must address the complication that the marginal constraint set $\mathcal{Q}_X$, and thus also the augmented cost function $\tilde{c}$ being convexified, depend on the observed sample through the empirical constraint. This dependence may, in principle, alter the optimal sample decomposition as the sample varies. The proof, in Appendix B.5, relies on the observation that the marginal constraint set $\mathcal{Q}_X$ is locally invariant to sample perturbations.

3.2.5 Simplicity

When the decomposition constraint is separable, the number of latent types employed in the constrained-optimal latent representation is bounded—as in Corollary 3 for the unconstrained case, albeit with a sharper bound.[20]

Corollary 7. *When the decomposition constraint $\mathcal{Q}$ is separable with marginal constraint $\mathcal{Q}_X$, then there exists a constrained-optimal latent representation that employs at most d+1 latent types, where d is the affine dimension of the marginal constraint set $\mathcal{Q}_X$.*

An analyst observing a high-dimensional sample and assuming a rich type space may find exact maximum likelihood mixture estimation computationally infeasible. To reduce dimensionality, the analyst may opt for a coarse sample decomposition with a low number of latent types, as in Bonhomme et al. (2022). Corollary 7 suggests an alternative procedure. The analyst can select a low-dimensional marginal constraint set $\mathcal{Q}_X$ such that the observed sample $q_0(x)$ lies within the convex hull of $\mathcal{Q}_X$, and then solve the constrained sample-decomposition problem. Corollary 7 guarantees that a low-dimensional optimal latent representation will emerge. A key advantage of this dimensionality reduction approach is its robustness: any alternative sample distribution within the convex hull of the original subsamples will yield the same set of latent types.

[20] This result again follows directly from Carathéodory's theorem. Since q_0 lies in the convex hull of optimal subsamples, which is at most d-dimensional, it can be expressed as a convex combination of at most $d+1$ elements of this set.

The following example illustrates how dimensionality is reduced when a boundedly rational agent relies on intuitive heuristics.

Example 2.3 (Dimensionality Reduction)**.** Recall the human resources manager who suffers from correlation neglect—captured by the set of primitive models $\mathcal{P}_X = \Delta(X_1)\times\Delta(X_2)$. The manager seeks to approximate a sample $q_0(x_1, x_2)$ over education (x_1) and experience (x_2) that exhibits correlation. Recall from Example 2.2 that in the absence of a decomposition constraint, the manager exactly fits the sample by decomposing it into degenerate atomic distributions, using a high-dimensional latent representation employing $|X_1|\times|X_2|$ latent types.

Now, assume a DAG-based decomposition constraint that restricts the manager to decomposing the sample using only a low-dimensional correlate of the data of interest. For each employee in her sample, the manager observes a triple (x_0, x_1, x_2), where the correlate—the spoiler variable x_0—stands for the employee's socioeconomic characteristic. The manager decomposes the sample solely based on socioeconomic characteristic, thereby imposing the conditional independence constraint $z \perp\!\!\!\perp_q (x_1, x_2) \mid x_0$. This is equivalent to the DAG-based constraint with the DAG

$$z \longrightarrow x_0 \longrightarrow x_1 \longrightarrow x_2 \qquad (14)$$

that generates the factorization constraint $q(x, z) = q(z)q(x_0 \mid z)q(x_1, x_2 \mid x_0)$. In this case, the marginal constraint set is given by

$$\mathcal{Q}_X = \{q(x_0, x_1, x_2) : q(x_1, x_2 \mid x_0) = q_0(x_1, x_2 \mid x_0)\}.$$

Since the dimensionality of this marginal constraint set is at most $|X_0|-1$, Corollary 7 implies that the constrained manager employs at most $|X_0|$ latent types. Specifically, when the classification of socioeconomic characteristic is coarse—resulting in $|X_0| \ll |X_1| \times |X_2|$—the sample decomposition becomes significantly constrained and the manager can no longer perfectly fit the observed sample. See Table 1 for an illustration. ▲

Unlike in the unconstrained case, the constrained-optimal latent representation need not be parsimonious. That is, there may exist distinct latent types $z \neq z'$ such that distinguishing between them has no implication for prediction, i.e. $p(x \mid z) = p(x \mid z')$, yet the associated subsamples differ: $q(x \mid z) \neq q(x \mid z')$.

low

	$x_2 = 0$	$x_2 = 1$
$x_1 = 0$	0.3	0.2
$x_1 = 1$	0.2	0.3

high

	$x_2 = 0$	$x_2 = 1$
0	0	0.2
1	0.2	0.6

Table 1: The sample consists of equal shares of individuals with low and high socioeconomic characteristic (x_0), with the empirical joint distributions of education (x_1) and experience (x_2) for each group shown in the tables above. The human resources manager, constrained by the DAG in (14), decomposes the sample by classifying employees solely on the basis of their socioeconomic characteristic, thereby forming subsamples with joint distributions $q(x_1, x_2 \mid z)$ from the convex hull of the two tables. Due to correlation neglect, the manager convexifies misspecification cost equal to mutual information between x_1 and x_2 over this convex hull; see Example 2.1. Since $\mathcal{Q}_X$ is one-dimensional, Corollary 7 implies that she optimally decomposes the sample into at most two subsamples. One subsample, comprising approximately 77% of the employees, pools individuals with high and low socioeconomic characteristic in proportions of roughly 0.65 and 0.35, respectively. The second subsample consists solely of individuals with low socioeconomic characteristic. Intuitively, this decomposition allows the manager to achieve a low correlation between x_1 and x_2 within the pooled subsample, thereby enabling a near-perfect fit for a large share of the observed data.

Example 4 (Parsimony Failure)**.** The observable variable $x = (x_1, x_2)$ takes values in $\{0,1\}^2$. The analyst is endowed with two primitive models, $\mathcal{P}_X = \{p_c, p_a\}$, where

$$\begin{aligned} p_c &= \left(\frac{1}{2}, 0, 0, \frac{1}{2}\right) \\ p_a &= \left(0, \frac{1}{2}, \frac{1}{2}, 0\right), \end{aligned}$$

with the tuples specifying probabilities over $(x_1, x_2) = 00, 01, 10, 11$. The subscript c denotes a correlated primitive model, while a denotes an anticorrelated one. The sample $q_0(x_1, x_2)$ is uniform over $\{0,1\}^2$. The sample decomposition is constrained by the DAG $x_1 \leftarrow z \rightarrow x_2$, which imposes the marginal constraint $\mathcal{Q}_X = \Delta(X_1) \times \Delta(X_2)$. Observe that $\mathcal{Q}_X$ is not convex.

The analyst optimally decomposes the sample into four degenerate subsamples, each assigning all probability mass to one of the four values $x = 00, 01, 10, 11$.[21] Let us label these subsamples by latent values $z = 00, 01, 10, 11$,

[21] To see that this is the optimal sample decomposition, note that any subsample containing

respectively. The analyst fits subsamples $z = 00$ and 11 using the correlated primitive model $p(x \mid z) = p_c(x)$ and fits subsamples for $z = 01$ and 10 using $p(x \mid z) = p_a(x)$. Thus the distinction between $z = 00$ and $z = 11$ is informative under the sample decomposition q, yet it is uninformative under the mixture model p (and similarly for the distinction between $z = 01$ and $z = 10$), generating parsimony failure. ▲

The next result shows that the parsimony failure in the previous example is driven by the non-convexity of the marginal constraint set. The proof mirrors that of parsimony in the unconstrained case, as stated in Corollary 4. When the marginal constraint set $\mathcal{Q}_X$ is convex, any non-parsimonious latent representation featuring distinct z and z' such that $p(x \mid z) = p(x \mid z')$ and $q(x \mid z) \neq q(x \mid z')$ can be improved upon by pooling the two subsamples.

Corollary 8. *When the marginal constraint set $\mathcal{Q}_X$ is convex, the constrained-optimal latent representation is parsimonious. Then, it employs at most $|\mathcal{P}_X|$ latent types: $|\mathrm{supp}(q(z))| = |\mathrm{supp}(p(z))| \leq |\mathcal{P}_X|$.*

For instance, the constrained-optimal latent representation under the statistical parity constraint from Example 1.4 is parsimonious, since the marginal constraint set $\mathcal{Q}_X = \{q(x_0, x_1) : q(x_0) = q_0(x_0)\}$ is convex. Hence, the number of latent types employed is at most $|\mathcal{P}_X| = 2$ in this case.

4 Literature

While economic theorists conveniently rely on exogenously specified priors over latent variables, typically referred to as types in a theoretical model, their empirical colleagues (and our analyst) proceed without such structure. Instead, the typical structural estimation uncovers the distribution of latent variables alongside other model parameters; see, for example, Heckman and Singer (1984), Kasahara and Shimotsu (2009), and Bonhomme et al. (2022). For instance, the stochastic choice literature infers heterogeneous latent individual preferences from observed choices, as in the mixed logit model of McFadden and Train (2000).

Our theoretical framework aligns with the empirical Bayes approach stemming from Robbins (1956). Under this approach, the analyst first estimates a

data points in both $\{00, 11\}$ and $\{01, 10\}$ incurs infinite misspecification cost. Given that the subsamples are restricted to independence, a fully informative sample decomposition is required to achieve a finite misspecification cost.

"prior" over the latent variable and subsequently updates beliefs in a Bayesian manner. Some of the theoretical models following this approach include Ortoleva (2012), Schwartzstein and Sunderam (2021), and Aina (2021). Their settings feature a fixed state and small samples, while we focus on multiple states and large samples.

Lindsay (1983) studies the structural properties of mixture models using a geometric approach that is related to, but distinct from, ours. His results overlap with ours in bounding the number of latent processes employed by the maximum likelihood mixture estimator. Relative to Lindsay, we extend this bound to settings in which the likelihood evaluation is subject to frictions, and we further analyze the sample decomposition, deriving its robustness to perturbations in the observed data.

Following the theoretical literature on misspecified learning, our analyst forms beliefs as a statistician, imperfectly approximating the empirical distribution. This literature, stemming from asymptotic results on misspecified learning by Berk (1966) and White (1982) and energized by Esponda and Pouzo (2016), endogenizes the data-generating process. We treat the data-generating process as exogenous, obviating the need to explicitly model the choice stage. We focus instead on the implications of misspecified learning for beliefs about latent variables. In an approach complementary to ours, Frick et al. (2024) study asymptotic posteriors about a fixed latent state under misspecified beliefs about the signal-generating process. In contrast, we examine beliefs over many i.i.d. latent states in a setting without a prior.

The estimation of mixture models often involves decomposing observations into segments attributed to distinct latent types. The pioneering Expectation–Maximization (E-M) algorithm of Dempster et al. (1977) iterates between fitting a mixture model to a segmented sample and resegmenting the sample based on the model's current fit. Our optimal latent representation corresponds to a fixed point of this procedure.

The E-M framework has had a significant impact on machine learning, where it led to Variational Bayesian methods of Jordan et al. (1999) and the variational autoencoder algorithm of Kingma and Welling (2013). Variational Bayesian methods approximate Bayesian updating as a constrained optimization problem, whereas variational autoencoders approximate the maximum likelihood estimation in mixture models. We study counterparts of these techniques in Section 3, where we impose frictions on the analyst's ability to evaluate likelihoods. Our constrained sample-decomposition corresponds to approximate

updating in the spirit of Variational Bayesian methods, while our constrained model-fitting parallels the role of variational autoencoders.[22] Our model p is referred to as the generative model while our sample decomposition q is called the recognition model in these literatures.

Our sample-decomposition constraint is a friction imposed on the evaluation of models' fits, which is complementary to the misspecification friction that restricts the choice of the hypothesized models. The likelihood-evaluation frictions remain underexplored in economic theory, with an exception of Aridor et al. (2020) and Aridor et al. (2025) who apply variational autoencoders to study bounded rationality in human decision-making.

While our sample decomposition constraint resembles constraints imposed on updating and estimation in machine learning, our motivation for studying such constraints is distinct. The machine learning literature primarily aims to optimize the performance of updating and estimation algorithms, while we seek to explain deviations of constrained analysts from rational benchmarks.

Our application of the concavification approach establishes technical connections to the Bayesian persuasion (Kamenica and Gentzkow, 2011) and rational inattention (Caplin and Dean, 2015) literatures, where these techniques are also employed. We develop a class of posterior-separable constraints that restrict information structures—which take the form of sample decompositions in our context—while preserving the applicability of concavification. This class of constraints may also prove relevant in persuasion and inattention settings. Specifically, we show that concavification techniques can accommodate a class of constraints naturally represented by causal networks, in the spirit of Spiegler (2016).

5 Summary

We study how economic agents represent data using latent constructs (e.g., preference types) without relying on exogenously specified priors. To this end, we adopt an empirical Bayes approach in which the prior over types and the meaning of a type are estimated via maximum likelihood from the sample itself. Our main result establishes that this estimation is equivalent to optimally decompos-

[22] See Ortoleva (2024) for a recent review of theoretical approaches to non-Bayesian updating within economics. The belief-updating framework of Dominiak et al. (2023) is closely related to Variational Bayesian methods, in that both represent updating as a constrained optimization problem with information-theoretic objective functions.

ing the sample into subsamples, each assigned to the primitive type minimizing its lack of fit, with the estimated prior corresponding to the relative sizes of these subsamples. This equivalence implies that optimal latent representations are structurally simple and locally robust to perturbations in the data. These properties extend to analysts facing frictions in likelihood evaluation, modeled as constraints on feasible sample decompositions, particularly for the class of separable constraints derived from causal directed acyclic graphs. We view the robustness of the types in our empirical Bayes approach as consistent with the common practice of fixing types exogenously in a theoretical model, even while performing comparative statics with respect to the distribution of types.

We have modeled the agent's belief formation which presumably informs subsequent decision-making. In contrast to Berk-Nash equilibrium models, we treat observations as unaffected by the agent's actions, thereby allowing belief formation to be analyzed independently of behavior. Future work could usefully incorporate interdependencies between actions and observations in strategic settings.

A Microfoundations For Constrained Likelihood

We first specify how a constrained analyst evaluates likelihood of any given mixture model and then allow her to select the mixture model that maximizes this constrained likelihood.

To formalize the analyst's inference process, we consider the limit of an arbitrarily large sample. Specifically, we examine a sequence of settings indexed by $n \in \mathbb{N}$. For each n, the analyst observes a sample $x^n = (x_i)_{i=1}^n$ with empirical distribution $q_0^n(x)$. The analyst reasons about the *extended* sample $(x_i, z_i)_{i=1}^n$ and its joint distribution $q(x, z)$ of the observable and latent variables. The analyst is endowed with a set $\mathcal{Q}^n \subseteq \Delta(X \times Z)$ of the joint distributions she considers, where X and Z are finite sets. Each distribution $q(x, z) \in \mathcal{Q}^n$ satisfies three types of constraints: (i) an integer constraint $q(x, z)n \in \mathbb{N}$, (ii) the empirical constraint $q(x) = q_0^n(x)$, and (iii) additional restrictions—arising, for example, from cognitive limitations or regulatory rules—that exclude certain joint distributions. Let S^n denote the set of extended samples whose empirical distribution lies in $\mathcal{Q}^n$. The analyst computes the *constrained likelihood* L^n of the observed sample x^n under mixture model p as the sum of the likelihoods of

all compatible extended samples:

$$L^n = \sum_{(x_i,z_i)_{i=1}^n \in S^n} \prod_{i=1}^n p(x_i, z_i).$$

To ensure a meaningful limit as $n \to \infty$, we assume that the sequence of constraint sets $\mathcal{Q}^n$ approximates the decomposition constraint $\mathcal{Q}$ from (7). To formalize this, let $\theta \in [0,1]$, define $\mathcal{Q}(0) = \mathcal{Q}$, and let $\mathcal{Q}(\theta) = \mathcal{Q}^{\lfloor 1/\theta \rfloor}$ denote the feasible set for $n = \lfloor 1/\theta \rfloor$. We assume that the correspondence $\mathcal{Q}(\theta)$ is continuous at $\theta = 0$—that is, both upper and lower hemicontinuous at that point.

The next result establishes that the constrained divergence $\mathrm{D}_{\mathcal{Q}}\big(p(x,z)\big)$, defined in (9), is equivalent—up to a trivial transformation—to the constrained likelihood evaluation of the model $p(x,z)$.

Lemma 2. *The constrained likelihood satisfies:*

$$\lim_{n\to\infty} \frac{1}{n} \ln L^n = -\mathrm{D}_{\mathcal{Q}}\big(p(x,z)\big) - \mathrm{H}\big(q_0(x)\big). \tag{15}$$

The lemma follows from Sanov's theorem. Given the constraint $\mathcal{Q}$, the analyst evaluates the likelihood of the model $p(x,z)$ as the probability that the empirical distribution of the extended sample $(x_i,z_i)_i$ drawn from $p(x,z)$ lies in $\mathcal{Q}$. In cases, where $p(x,z) \notin \mathcal{Q}$, this corresponds to evaluating the probability of an atypical event. Sanov's theorem implies that the normalized log-likelihood of this atypical event converges to the negative of the divergence $\mathrm{KL}\big(\tilde{q}(x,z) \| p(x,z)\big)$ minimized over $\tilde{q}(x,z) \in \mathcal{Q}$, giving (15) (up to the constant term $\mathrm{H}\big(q_0(x)\big)$, which appears because the marginal distribution $q(x) = q_0(x)$ is fixed). For completeness, we now provide a proof from first principles:

Proof of Lemma 2. The p-likelihood of a single extended sample $(x_i,z_i)_{i=1}^n$ with empirical distribution $\tilde{q}(x,z)$ is given by

$$\prod_{i=1}^n p(x_i,z_i) = \prod_{x,z} p(x,z)^{\tilde{q}(x,z)n}.$$

Accounting for their number, the p-likelihood of all extended samples with distribution $\tilde{q}(x,z)$ is

$$\ell^n(\tilde{q}) := \mathcal{N}_n(\tilde{q}) \prod_{x,z} p(x,z)^{\tilde{q}(x,z)n}, \tag{16}$$

where $\mathcal{N}_n(\tilde{q})$ denotes the number of distinct extended samples $(x_i, z_i)_{i=1}^n$ that have the empirical distribution $\tilde{q}(x, z)$ and match the observed sample x^n on the margin. Note that the constrained likelihood is given by

$$L^n = \sum_{\tilde{q} \in \mathcal{Q}^n} \ell^n(\tilde{q}).$$

In the first step, we approximate $\frac{1}{n} \ln \ell^n(\tilde{q})$. For this, observe that

$$\mathcal{N}_n(\tilde{q}) = \prod_x \mathcal{N}'_{\tilde{q}(x)n}\big(\tilde{q}(z \mid x)\big),$$

where $\mathcal{N}'_m\big(\pi(z)\big)$ is the number of the distinct sequences $(z_1, \ldots, z_m)$ of the length m with the empirical distribution $\pi(z)$. This holds because to compute $\mathcal{N}_n(\tilde{q})$ we can, for each value x, consider a subsequence $(i_k)_k$ of length $\tilde{q}(x)n$ such that $x_{i_k} = x$ for all k and the empirical distribution of z_{i_k} is $\tilde{q}(z \mid x)$. Then, $\mathcal{N}'_{\tilde{q}(x)n}\big(\tilde{q}(z \mid x)\big)$ is the number of distinct permutations for each such subsequence.

Theorem 11.1.3 in Cover and Thomas (1999) provides the following bounds:

$$\frac{1}{(m+1)^{|Z|}} \exp\big[m \times \mathrm{H}\big(\pi(z)\big)\big] \leq \mathcal{N}'_m\big(\pi(z)\big) \leq \exp\big[m \times \mathrm{H}\big(\pi(z)\big)\big].$$

Substituting these bounds into (16) for each x, with $m = \tilde{q}(x)n$, and using the fact that $\tilde{q}(x)n \leq n$, taking logarithm, and normalizing by n gives bounds

$$\begin{aligned} &\mathrm{E}_{\tilde{q}(x,z)} \ln p(x, z) + \sum_x \tilde{q}(x)\, \mathrm{H}\big(\tilde{q}(z \mid x)\big) - |Z| \times |X| \frac{\ln(n+1)}{n} \\ &\leq \ \frac{1}{n} \ln \ell^n(\tilde{q}) \ \leq \\ &\mathrm{E}_{\tilde{q}(x,z)} \ln p(x, z) + \sum_x \tilde{q}(x)\, \mathrm{H}\big(\tilde{q}(z \mid x)\big). \end{aligned} \tag{17}$$

In the second step, we use the bounds from (17) to derive bounds for the constrained likelihood $L^n = \sum_{\tilde{q} \in \mathcal{Q}^n} \ell^n(\tilde{q})$. For this, we define value

$$v(\mathcal{Q}^n) = \max_{\tilde{q}(x,z) \in \mathcal{Q}^n} \mathrm{E}_{\tilde{q}(x,z)} \ln p(x, z) + \sum_x \tilde{q}(x)\, \mathrm{H}\big(\tilde{q}(z \mid x)\big),$$

and observe that

$$\begin{aligned} & v(\mathcal{Q}^n) - |Z| \times |X| \frac{\ln(n+1)}{n} \\ \leq\ & \frac{1}{n} \ln L^n \ \leq \\ & v(\mathcal{Q}^n) + |Z| \times |X| \times \frac{\ln(n+1)}{n}. \end{aligned}$$

For the upper bound, we used that for a given n, distributions $\tilde{q}(x,z)$ take values in $\{\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n}{n}\}$ and thus $|\mathcal{Q}^n| \leq (n+1)^{|Z|\times|X|}$, and, additionally, $\frac{1}{n} \ln \ell^n(\tilde{q}) \leq v(\mathcal{Q}^n)$ for all $\tilde{q} \in \mathcal{Q}^n$. Thus, $L^n \leq \exp\big(nv(\mathcal{Q}^n)\big)(n+1)^{|Z|\times|X|}$. Taking logarithm and normalizing by n gives the upper bound. For the lower bound, we use the fact that $L^n \geq \max_{\tilde{q}\in\mathcal{Q}^n} \ell^n(\tilde{q})$.

In the third step, we observe that the Maximum Theorem, together with the fact that $\frac{\ln(n+1)}{n} \to 0$, imply that $\frac{1}{n} \ln L^n \to v(\mathcal{Q})$.

Finally,

$$\mathrm{E}_{\tilde{q}(x,z)} \ln p(x,z) + \sum_x \tilde{q}(x)\, \mathrm{H}\big(\tilde{q}(z \mid x)\big) + \mathrm{H}\big(q_0(x)\big) = -\,\mathrm{KL}\big(\tilde{q}(x,z) \| p(x,z)\big),$$

implying equation (15). ■

We now assume that the analyst selects a mixture model to maximize the constrained likelihood—that is, she minimizes the constrained divergence. Accordingly, we introduce a *constrained* model-fitting problem as the natural counterpart to the model-fitting problem (1):[23]

$$\begin{aligned} \min_{\tilde{p}(x,z)} \quad & \mathrm{D}_{\mathcal{Q}}\big(\tilde{p}(x,z)\big) \\ \text{s.t.} \quad & \tilde{p}(x \mid z) \in \mathcal{P}_X \text{ for all } z \in \operatorname{supp}\big(\tilde{p}(z)\big). \end{aligned} \tag{18}$$

The equivalence between the model-fitting and sample-decomposition problems extends to their constrained counterparts:

Proposition 3 (Constrained Decomposition Equivalence)**.** *The constrained model-fitting problem* (18) *is equivalent to the constrained sample-decomposition*

[23] For existence, we assume that there exists at least one feasible pair $p(x,z)$, $q(x,z)$ such that the support of $q(x,z)$ is a subset of the support of $p(x,z)$. This pair attains a finite objective value, and the set of feasible model pairs achieving at most this value is compact. Continuity of the objective then ensures that a solution exists.

problem (8). *Specifically, the mixture model* $p(x,z) = p(z)p(x \mid z)$ *solves problem* (18) *if and only if*

$$\begin{aligned} p(z) &= q(z), \text{ and} \\ p(x \mid z) &\in \arg\max_{\tilde{p}(x) \in \mathcal{P}_X} \mathrm{E}_{q(x|z)} \ln \tilde{p}(x) \quad \forall z \in \operatorname{supp}(q(z)), \end{aligned}$$

where $q(x,z)$ *solves the constrained sample-decomposition problem* (8).

The proof is analogous to that of Proposition 1:

Proof of Proposition 3. Recalling the definition of the constrained divergence $\mathrm{D}_{\mathcal{Q}}(p(x,z))$, the constrained model-fitting problem (18) is equivalent to

$$\begin{aligned} \min_{\tilde{p}(x,z), \tilde{q}(x,z)} \quad & \mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) \\ \text{s.t.} \quad & \tilde{p}(x \mid z) \in \mathcal{P}_X \quad \forall z \in \operatorname{supp}\left(\tilde{p}(z)\right), \\ & \tilde{q}(x,z) \in \mathcal{Q}. \end{aligned} \tag{19}$$

Using the chain rule for the Kullback-Leibler divergence, the objective can be rewritten as

$$\mathrm{KL}\left(\tilde{q}(x,z) \parallel \tilde{p}(x,z)\right) = \mathrm{KL}\left(\tilde{q}(z) \parallel \tilde{p}(z)\right) + \mathrm{E}_{\tilde{q}(z)} \mathrm{KL}\left(\tilde{q}(x \mid z) \parallel \tilde{p}(x \mid z)\right).$$

Fixing any given sample decomposition $\tilde{q}(x,z)$, the inner minimization with respect to $\tilde{p}(x,z)$ implies that the minimizer p satisfies

$$p(z) = \tilde{q}(z) \quad \text{and} \quad p(x \mid z) \in \arg\min_{\tilde{p} \in \mathcal{P}_X} \mathrm{KL}\left(\tilde{q}(x \mid z) \parallel \tilde{p}\right) = \arg\max_{\tilde{p} \in \mathcal{P}_X} \mathrm{E}_{\tilde{q}(x|z)} \ln \tilde{p}(x),$$

for all z in the support of $p(z) = \tilde{q}(z)$

Substituting these optimizers into the objective yields $\mathrm{E}_{\tilde{q}(z)}\, c\big(\tilde{q}(x \mid z)\big)$. The outer minimization with respect to $\tilde{q}(x,z) \in \mathcal{Q}$ then corresponds exactly to the constrained sample-decomposition problem (8), completing the proof. ■

B Proofs

B.1 Proof of Proposition 1

Recall that the model-fitting problem (1) is equivalent to

$$\begin{aligned} \min_{\tilde{p}(x,z)} \quad & \mathrm{E}_{q_0(x)} \mathrm{KL}\left(q_0(x) \| \tilde{p}(x)\right) \\ \text{s.t.} \quad & \tilde{p}(x \mid z) \in \mathcal{P}_X \text{ for all } z \in \operatorname{supp}(\tilde{p}(z)). \end{aligned}$$

Mixture model $p(x, z)$ solves this problem if and only if $p(x, z)$ and $q(x, z) = q_0(x)p(z \mid x)$ jointly solve:

$$\begin{aligned} \min_{\tilde{p}(x,z),\tilde{q}(x,z)} \quad & \mathrm{KL}\left(\tilde{q}(x, z) \parallel \tilde{p}(x, z)\right) \\ \text{s.t.} \quad & \tilde{p}(x \mid z) \in \mathcal{P}_X \quad \forall z \in \operatorname{supp}\left(p(z)\right), \\ & \tilde{q}(x) = q_0(x). \end{aligned} \tag{20}$$

This equivalence holds because, by applying the chain rule for the Kullback-Leibler divergence to the objective in (20), we obtain

$$\mathrm{KL}\left(\tilde{q}(x, z) \parallel \tilde{p}(x, z)\right) = \mathrm{KL}\left(q_0(x) \parallel \tilde{p}(x)\right) + \sum_{x \in X} q_0(x) \mathrm{KL}\left(\tilde{q}(z \mid x) \parallel \tilde{p}(z \mid x)\right),$$

and minimization of each $\mathrm{KL}\left(\tilde{q}(z \mid x) \parallel \tilde{p}(z \mid x)\right)$ term with respect to $\tilde{q}(z \mid x) \in \Delta(Z)$ eliminates the second term in the displayed expression by setting $\tilde{q}(z \mid x) = \tilde{p}(z \mid x)$ for all x (this implies the statement $q(z \mid x) = p(z \mid x)$ in Corollary 1.)

We again apply the chain rule to the objective in (20) to write

$$\mathrm{KL}\left(\tilde{q}(x, z) \parallel \tilde{p}(x, z)\right) = \mathrm{KL}\left(\tilde{q}(z) \| \tilde{p}(z)\right) + \sum_{z \in \operatorname{supp}(\tilde{q}(z))} \tilde{q}(z) \mathrm{KL}\left(\tilde{q}(x \mid z) \| \tilde{p}(x \mid z)\right).$$

Optimization over $\tilde{p}(z) \in \Delta(Z)$ implies $p(z) = \tilde{q}(z)$ and eliminates the first term on the right (this implies the statement $q(z) = p(z)$ in Corollary 1.)

Further optimization over $\tilde{q}(x, z)$ and $\left(\tilde{p}(x \mid z)\right)_z$ implies that when $p(x, z)$ and $q(x, z)$ solve (20), then $q(x, z)$ is an optimal decomposition that solves the sample-decomposition problem (2) and, for each $z \in \operatorname{supp}\left(q(z)\right)$, $p(x \mid z)$ minimizes $\mathrm{KL}\left(q(x \mid z) \| \tilde{p}(x \mid z)\right)$ within $\mathcal{P}_X$, and thus $p(x \mid z)$ maximizes expected log-likelihood for given $q(x \mid z)$, as stated in (5). ■

B.2 Proof of Corollary 2

If $p(x, z)$ and $q(x, z)$ constitute an optimal latent representation for sample $q_0(x)$, then $q(x \mid z)$ are tangency points of the cost function $c(\cdot; \mathcal{P}_X)$ and the hyperplane tangent to its convex envelope $C(\cdot)$ at $q_0(x)$. For any sample $q_0'(x) = \sum r(z)q(x \mid z)$, the subsamples $q(x \mid z)$ remain tangency points of the convexification and $r(z)$ is their distribution. Thus, $q'(x, z) = r(z)q(x \mid z)$ solves the sample-decomposition problem (2) and the mixture model $p'(x, z) = r(z)p(x \mid z)$ satisfies the requirements (4) and (5). It follows that (p', q') form an optimal latent representation for the new sample $q_0'(x)$. ■

B.3 Proof of Corollary 4

When defining sample decompositions, we have assumed without loss of generality that $q(x \mid z) \neq q(x \mid z')$ for all distinct pairs z, z'. Therefore, to establish parsimony, it suffices to show that $p(x \mid z) \neq p(x \mid z')$.

Assume for contradiction $p(x \mid z) = p(x \mid z')$ for some distinct pair z, z'. Construct an alternative latent representation (p', q') by merging z and z' into a single latent value z'', and leaving all other values unchanged. Specifically, define:

- $q'(x, \hat{z}) = q(x, \hat{z})$, $p'(x, \hat{z}) = p(x, \hat{z})$ for all $\hat{z} \neq z, z'$,
- $q'(z'') = p'(z'') = q(z) + q(z') = p(z) + p(z')$,
- $q'(x \mid z'') = \frac{q(z)q(x|z)+q(z')q(x|z')}{q(z)+q(z')}$,
- $p'(x \mid z'') = p(x \mid z) = p(x \mid z')$.

This merged latent representation improves the objective in the sample-decomposition problem (2), because:

$$\begin{aligned} q(z)c\big(q(x \mid z)\big) + q(z')c\big(q(x \mid z')\big) &= q(z)\,\mathrm{KL}\left(q(x \mid z) \| p(x \mid z)\right) + q(z')\,\mathrm{KL}\left(q(x \mid z') \| p(x \mid z)\right) \\ &> q'(z'')\,\mathrm{KL}\left(q'(x \mid z'') \| p(x \mid z)\right) \\ &\geq q'(z'')c\big(q'(x \mid z'')\big), \end{aligned}$$

where the strict inequality follows from the strict convexity of KL divergence in its first argument and the last inequality uses the definition of the cost function $c(\cdot)$. ■

B.4 Proof of Lemma 1

We aim to prove the equivalence:

$$q(x,z) \in \mathcal{Q} \quad \Leftrightarrow \quad q(x \mid z) \in \mathcal{Q}_X \text{ for all } z \in \operatorname{supp}\big(q(z)\big) \text{ and } q(x) = q_0(x).$$

"⇐" Direction. This direction is immediate. Suppose $q(x \mid z) \in \mathcal{Q}_X$ for all $z \in \operatorname{supp}\big(q(z)\big)$ and $q(z) \in \Delta(Z)$ is arbitrary. Then the joint distribution $q(x,z) = q(z)q(x \mid z)$ satisfies the DAG-based factorization required in (13), and the marginal constraint $q(x) = q_0(x)$ is assumed. Hence, $q(x,z) \in \mathcal{Q}$.

"⇒" Direction. Suppose $q(x,z) \in \mathcal{Q}$. Then by definition, $q(x,z)$ satisfies the DAG factorization in (13), and $q(x) = q_0(x)$. We must show that for all $z \in \operatorname{supp}\big(q(z)\big)$, the conditional distribution $q(x \mid z) \in \mathcal{Q}_X$, where $\mathcal{Q}_X$ is defined in Lemma 1.

First, for each fixed z, the conditional $q(x \mid z)$ inherits the DAG factorization structure from the global factorization of $q(x,z)$ (i.e., satisfies point (i) in the definition of $\mathcal{Q}_X$.)

Second, consider any variable x_k such that $z \notin \operatorname{pa}(x_k)$. Since $q(x,z) \in \mathcal{Q}$, the DAG factorization constraint implies that

$$q\big(x_k \mid \operatorname{pa}(x_k), z\big) = q\big(x_k \mid \operatorname{pa}(x_k)\big),$$

i.e., x_k is independent of z, conditional on its DAG parents. Moreover, since the empirical constraint requires

$$\mathrm{E}_{q(z)}\, q\big(x_k \mid \operatorname{pa}(x_k), z\big) = q_0\big(x_k \mid \operatorname{pa}(x_k)\big),$$

it must be that

$$q\big(x_k \mid \operatorname{pa}(x_k), z\big) = q_0\big(x_k \mid \operatorname{pa}(x_k)\big) \quad \forall z \in \operatorname{supp}\big(q(z)\big),$$

as required for point (ii) in the definition of $\mathcal{Q}_X$. ■

B.5 Proof of Proposition 2

The marginal constraint set $\mathcal{Q}_X(q_0)$ depends on the sample $q_0(x)$, as indicated here by the explicit argument. We first observe a local invariance property: for

any $q_0'(x) \in \text{co}\big(\mathcal{Q}_X(q_0)\big)$, it holds that

$$\mathcal{Q}_X(q_0') = \mathcal{Q}_X(q_0).$$

To verify this, recall from Lemma 1 that the dependence of $\mathcal{Q}_X(q_0)$ on q_0 arises only through point (ii) of its definition. Specifically, for each variable x_k such that $z \notin \text{pa}(x_k)$, point (ii) imposes the empirical constraint:

$$q(x_k \mid \text{pa}(x_k)) = q_0(x_k \mid \text{pa}(x_k)),$$

for all $q(x) \in \mathcal{Q}_X$. Now, suppose $q_0' \in \text{co}\big(\mathcal{Q}_X(q_0)\big)$. Since all elements of $\mathcal{Q}_X(q_0)$ agree on the conditional distributions $q(x_k \mid \text{pa}(x_k))$ for such x_k, any convex combination q_0' must also satisfy:

$$q_0'(x_k \mid \text{pa}(x_k)) = q_0(x_k \mid \text{pa}(x_k)).$$

Thus, the set $\mathcal{Q}_X(q_0')$ imposes the same conditional constraints and is identical to $\mathcal{Q}_X(q_0)$.

The proposition then follows from the fact that when the sample q_0' is in the convex hull of the optimal subsamples for the sample q_0, then optimal decomposition is for both samples q_0 and q_0' given by the convexification of the same augmented cost function $\tilde{c}_{q_0'}(\cdot) = \tilde{c}_{q_0}(\cdot)$. ■

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 auto-encoder approach. biorxiv preprint 2020–05, Cold Spring Harbor Laboratory.

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

Bervoets, S., M. Faure, and L. Renou (2025). Non-bayesian learning in misspecified models. *arXiv preprint arXiv:2503.18024*.

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.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806.

Cover, T. M. and J. A. Thomas (1999). *Elements of Information Theory*. John Wiley & Sons.

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, arXiv.

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 (2024). Welfare comparisons for biased learning. *American Economic Review 114*(6), 1612–1649.

He, K., F. Sandomirskiy, and O. Tamuz (2021). Private private information. *arXiv preprint arXiv:2112.14356*.

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.

Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic theory 123*(2), 81–104.

Jehiel, P. and G. Weber (2025). Endogenous clustering and analogy-based expectation equilibrium. *arXiv preprint arXiv:2505.13022*.

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.

Kamenica, E. and M. Gentzkow (2011). Bayesian persuasion. *American Economic Review 101*(6), 2590–2615.

Kasahara, H. and K. Shimotsu (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica 77*(1), 135–175.

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*. Forthcoming.

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.

Matyskova, L. and A. Montes (2023). Bayesian persuasion with costly information acquisition. *Journal of Economic Theory 211*, 105678.

McFadden, D. and K. Train (2000). Mixed MNL models for discrete response. *Journal of applied Econometrics 15*(5), 447–470.

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

Ortoleva, P. (2024). Alternatives to Bayesian updating. *Annual Review of Economics 16*, 545–579.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

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.

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

Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290.

Spiegler, R. (2020). Can agents with causal misperceptions be systematically fooled? *Journal of the European Economic Association 18*(2), 583–617.

Strack, P. and K. H. Yang (2024). Privacy-preserving signals. *Econometrica 92*(6), 1907–1938.

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