Deep Latent Factor Models for the Social Sciences

Germain Gauthier, Philine Widmer, Elliott Ash

March 25, 2026

Preliminary version. Comments welcome.

Abstract

We propose a general class of latent factor models for structured data (e.g., surveys, voting records) and unstructured data (e.g., text, images, and other high-dimensional objects). The framework is modular and allows each observational unit to be measured in multiple data types, with the latent factor capturing common variation across modalities and, when desired, predicting auxiliary outcomes. Our estimation approach is likelihood-based: we maximize an evidence lower bound (ELBO) on the marginal likelihood, with the bound parameterized by neural networks and optimized by stochastic gradient methods with amortized inference. Theoretically, we provide conditions under which the ELBO is tight and yields exact inference, and we clarify how approximation error decomposes into variational and amortization components. Empirically, the method is fast and scalable, performs well in Monte Carlo simulations, and accommodates missing modalities. We demonstrate its usefulness in three applications in economics and political economy: (1) decomposition of polarized rhetoric into topics versus framing; (2) interpretable macroeconomic forecasting with business news; and (3) joint text–image topic modeling of political communication on social media.[1]

Keywords: Deep learning, Latent factor models, Multimodal data

[1]Author details: Germain Gauthier (corresponding author), Bocconi University, germain.gauthier@unibocconi.it; Philine Widmer, Paris School of Economics, philine.widmer@psemail.eu; Elliott Ash, ETH Zurich, ashe@ethz.ch. We thank Stephen Hansen and Xavier d'Haultfoeuille for helpful feedback. We thank Anton Boltachka, Benjamin Kohler, Shimpei Kojio, Federica Kulka, Guanlan Liu, and Wei Pin for their excellent research assistance.

1 Introduction

Many empirical questions in economics and political economy require recovering a small number of latent variables from high-dimensional measurements. Classic examples include factor models for macroeconomic and financial time series and latent-trait models for political behavior. Increasingly, however, the relevant measurements are *unstructured* and *multimodal*: text, images, audio, and other high-dimensional objects are observed alongside (or instead of) structured outcomes, and different observations may be measured in different modalities. These settings pose two econometric challenges. First, once the measurement equations are allowed to be flexible and nonlinear, the marginal likelihood involves high-dimensional integrals that are typically intractable. Second, in multimodal datasets, researchers need a coherent notion of posterior inference for the latent factor when only a subset of modalities is observed.

This paper develops a likelihood-based econometric framework for latent factor models with multiple data modalities and flexible measurement equations. For each unit i, we posit a low-dimensional latent factor z_i that generates up to M observed modalities w_{im} and up to K auxiliary outcomes y_{ik}, conditional on covariates that may shift latent-factor prevalence, modality content, and outcome mappings. The model is modular in the sense that each modality can be equipped with a likelihood and function class appropriate to the data type (e.g., multinomial for text, Gaussian for continuous measurements, or more expressive decoders for complex objects). This specification nests standard latent-variable models such as probabilistic PCA (Tipping and Bishop, 1999) and topic-model variants with covariates and supervision (Mcauliffe and Blei, 2007) as special cases, while accommodating modern unstructured data.

Our estimation approach addresses the intractable likelihood by maximizing an evidence lower bound (ELBO) on the marginal log likelihood using amortized variational inference (Blei et al., 2017; Kingma, 2013; Kingma et al., 2019). Amortization replaces per-observation variational optimization with a shared inference function that maps observed data into a tractable variational family, enabling scalable estimation and out-of-sample inference. In multimodal settings, a key econometric issue is that the target posterior is defined for every subset of observed modalities. We therefore train under random modality subsampling and construct subset-specific variational posteriors by fusing unimodal encoders with a corrected product-of-experts rule. The correction removes redundant prior factors and mirrors the multiplicative structure of the true posterior implied by conditional independence, in contrast to mixture-based fusion rules (Wu and Goodman, 2018; Shi et al., 2019; Suzuki et al., 2016).

The central theoretical question is when ELBO-based procedures can be interpreted as exact likelihood-based inference, and when they should be viewed as approximation methods. We provide general sufficient conditions under which the amortized ELBO is uniformly tight on compact domains. The result clarifies how the inference error decomposes into a *variational gap* (from restricting the approximating family) and an *amortization gap* (from restricting attention to an encoder class) (Cremer et al., 2018), and it identifies conditions under which both components can be made arbitrarily small. We then study an extended conditional linear–Gaussian special case in

which the exact posterior is available in closed form. In this benchmark, (i) the corrected product-of-experts fusion rule is exact when unimodal encoders coincide with unimodal posteriors, and (ii) ELBO maximization inherits standard consistency and asymptotic normality results from classical likelihood-based M-estimation. This benchmark illustrates how, under correct specification and realizability of the posterior map, amortized variational estimation can recover familiar frequentist guarantees.

We implement the framework in a concrete instantiation tailored to text-as-data settings with rich metadata, which we refer to as the generalized topic model (GTM). GTM can be used for exploratory measurement and for prediction when auxiliary outcomes are observed. In contrast to two-step pipelines that estimate topics first and then fit a downstream predictive model, GTM allows joint estimation of latent factors and prediction targets within a single objective, permitting an explicit tradeoff between reconstruction fit and predictive accuracy (Kingma et al., 2014; Higgins et al., 2017). We illustrate the scope of the framework in three applications. First, using the *U.S. Congressional Record* (1873–2016), we measure speech polarization and decompose it into changes in topic choice versus within-topic framing, connecting the decomposition to phrase-based polarization measurement (Gentzkow et al., 2019). Second, using *Wall Street Journal* business news (1984–2012), we forecast a macroeconomic indicator from text and compare the joint topic–prediction approach to a standard LDA+LASSO pipeline, in the spirit of recent work that incorporates text-derived topics into macro-finance forecasting (Larsen and Thorsrud, 2019; Thorsrud, 2020; Bybee et al., 2024). Third, we estimate joint topics over text and images in U.S. Members of Congress' Facebook posts (2010–2022), using multimodal representations to enable coherent cross-modal latent factors (Zhang et al., 2024).

The paper's contribution is to provide (i) a modular econometric latent-factor specification that accommodates multimodal unstructured measurements, (ii) a scalable estimation strategy with principled treatment of missing modalities via corrected posterior fusion, and (iii) theoretical conditions clarifying when amortized variational inference delivers exact inference and when it yields an approximation, together with a tractable benchmark establishing classical asymptotics.

The remainder of the paper is organized as follows. Section 2 presents the model, estimation approach, and statistical guarantees. Section 3 reports some simulations and empirical validations. Sections 4–6 present the three applications. The appendix collects special-case mappings, derivations, and proofs.

2 Deep latent factor models

This section specifies our data-generating process and describes our estimation approach. The econometric object of interest is a low-dimensional latent factor z_i that summarizes common variation across possibly high-dimensional measurements and potentially can be used to predict auxiliary outcomes. The measurements can also stride multiple modalities: text, images, or other forms of "unstructured" data. Our goal is to estimate (i) a measurement model linking z_i to each observed modality and (ii) an approximation to the conditional distribution of z_i given the ob-

served data.

The main practical challenge is computational: the marginal likelihood typically involves an integral over the latent factors that has no closed form when likelihoods are nonconjugate or parameterized flexibly (e.g., by neural networks). We therefore use variational inference. Concretely, we derive an evidence lower bound (ELBO) on the log marginal likelihood and maximize it by stochastic gradient methods. To make inference feasible at large n, we use a shared (amortized) inference function—an encoder—that maps observed data to a tractable approximate posterior distribution for z_i.

2.1 Data-generating process (DGP)

2.1.1 The model

For each unit $i \in \{1,\dots,n\}$ we observe up to M response variables ("modalities") w_{im}, indexed by $m \in \{1,\dots,M\}$, up to K auxiliary outcomes y_{ik}, indexed by $k \in \{1,\dots,K\}$, and covariates x_i. The econometric object of interest is an L-dimensional latent factor z_i that captures the common variation in $(w_{i1},\dots,w_{iM})$ and is predictive for $(y_{i1},\dots,y_{iK})$.

We partition covariates into three sets to make their roles explicit. First, prevalence covariates $x_i^p \in \mathcal{X}^p$ shift the distribution of latent factors across units through the conditional distribution $p_\psi(z_i \mid x_i^p)$. In topic-model applications, for example, x_i^p can include metadata that shifts topic prevalence. Second, covariates $x_i^c \in \mathcal{X}^c$ shift the conditional distribution of the response variables w_{im} given z_i through $p_{\theta_m}(w_{im} \mid z_i, x_i^c)$. These covariates can be used to allow the mapping from latent factors to observables to vary with observed characteristics (e.g., to model systematic changes in how a topic is expressed). Third, covariates $x_i^s \in \mathcal{X}^s$ shift the conditional distribution of auxiliary outcomes y_{ik} given z_i through $p_{\theta_k}(y_{ik} \mid z_i, x_i^s)$. We use this partition because these three roles are conceptually distinct: covariates may shift who has which latent traits (x_i^p), how latent traits manifest in measurements (x_i^c), and how latent traits map into auxiliary outcomes (x_i^s).

Let $z_i \in \mathcal{Z} \subseteq \mathbb{R}^L$ denote an L-dimensional vector of latent factors for unit i. We assume that L is fixed and known. Throughout, we use the term *prior* for the distribution of z_i given x_i^p, but it should be understood as a modeling assumption rather than a subjective Bayesian prior. We assume *conditional independence* across modalities and outcomes given the latent factor and the relevant covariates:

$$
\begin{aligned}
&\textbf{Latent factors:} && z_i \sim p_\psi(z_i \mid x_i^p), \\
&\textbf{Response variables:} && w_{im} \sim p_{\theta_m}(w_{im} \mid z_i, x_i^c), \quad m = 1,\dots,M, \\
&\textbf{Auxiliary outcomes:} && y_{ik} \sim p_{\theta_k}(y_{ik} \mid z_i, x_i^s), \quad k = 1,\dots,K.
\end{aligned}
\tag{1}
$$

This formulation nests a variety of well-studied latent factor models as special cases (e.g., probabilistic PCA and several topic models). Appendix B maps a set of standard models into the notation of Eq. (1). At the same time, the specification is deliberately modular: it allows the researcher to choose likelihoods and function classes that are appropriate for each modality. For example,

p_{θ_m} may be Gaussian for continuous data or categorical/multinomial for discrete outcomes such as words. If the researcher wishes to model unstructured data types (e.g., images or video), p_{θ_m} can be parameterized by expressive functions such as deep neural networks.

2.1.2 Identifiability

Latent factor models are generally not identifiable—that is, different parameter configurations can produce observationally equivalent distributions over observables. This non-identifiability arises because latent factors are unobserved, leaving transformations of the latent space empirically indistinguishable. Of course, identifiability is a property of the full DGP one specifies, and in particular of the decoder/likelihood mapping from z_i to observables. As a result, if identifiability is required for a given application, it is in the researcher's hands to impose it through modeling restrictions (e.g., constraints on the decoder, structured priors, or auxiliary information) rather than to expect it to emerge from unconstrained training.

In the simplest linear case, probabilistic principal component analysis (PPCA) is identifiable only up to trivial reparametrizations: sign flips, location shifts, and scale transformations of the latent factors (Tipping and Bishop, 1999). These transformations do not change the likelihood but alter the representation of z_i. For instance, if the model places high probability on (z_i, θ), it places equal probability on $(-z_i, -\theta)$ or any rotation of the latent space. Researchers typically resolve these ambiguities by imposing normalizing constraints, such as zero mean and unit variance for the latent factors, and ordering conventions for the dimensions.

Deeper models face more severe identifiability challenges. When decoders are parameterized by neural networks, the set of non-identifiable transformations expands considerably. In particular, any diffeomorphism (smooth, invertible transformation) $\phi : \mathcal{Z} \to \mathcal{Z}$ applied to the latent space can be compensated by adjusting the decoder to yield an equivalent generative model. Formally, if $p_\theta(w_i \mid z_i)$ generates the data, then so does $p_{\theta'}(w_i \mid \phi(z_i))$ for appropriately transformed parameters θ' (Khemakhem et al., 2020). This means that, without additional assumptions, one can obtain infinitely many parameterizations that fit the data equally well but correspond to fundamentally different latent representations.

Identifiability can be achieved under specific structural assumptions. For instance, Khemakhem et al. (2020) shows that latent variables in nonlinear independent component analysis (ICA) can be identified up to permutation and component-wise transformations when the generative model satisfies certain conditional independence structures, and the data include auxiliary variables that modulate the latent distribution. Similarly, recent work on identifiable VAEs imposes constraints such as sparsity in the decoder (Locatello et al., 2019; Moran et al., 2021) or leverages temporal or distributional structure (Hyvarinen et al., 2019). However, such conditions are often restrictive and may not hold in typical social science applications, or may be more appropriate to model some data types than others (e.g., a sparse decoder may be relevant to model topics in text as data, but will likely underperform for visual data).

In practice, we observe that the use of regularization during training often acts as an implicit selection rule among the many observationally equivalent solutions. Regularization techniques

such as weight decay, dropout, or architectural choices like limiting network depth or width encourage simpler models. These inductive biases guide the optimization process toward solutions that are not only consistent with the data but also simpler in a sense defined by the regularization scheme. One can interpret this as the model selecting the "simplest route" to explain the data—a form of Occam's razor embedded in the learning algorithm. While this does not resolve the fundamental non-identifiability, it often yields interpretable and stable representations in practice. Researchers should nonetheless remain aware that the learned latent factors are one of many possible representations, and conclusions about their substantive meaning should be appropriately qualified.

2.2 Estimation via amortized variational inference

2.2.1 The marginal likelihood

Given the DGP in Eq. (1), the (conditional) marginal likelihood for a single unit i (with all M modalities observed) is

$$p_{\theta,\psi}(w_{i,1:M}, y_{i,1:K} \mid x_i) \;=\; \int p_\psi(z_i \mid x_i^p) \left[\prod_{m=1}^{M} p_{\theta_m}(w_{im} \mid z_i, x_i^c)\right] \left[\prod_{k=1}^{K} p_{\theta_k}(y_{ik} \mid z_i, x_i^s)\right] \mathrm{d}z_i. \tag{2}$$

When the decoder distributions are nonlinear functions of z_i (as is typical with deep learning parameterizations) and/or when likelihoods are nonconjugate, the integral in Eq. (2) has no closed form. Direct maximization of the marginal likelihood would then require repeated numerical integration and differentiation through the integral. In high-dimensional, multimodal settings this is typically infeasible: evaluating and differentiating the likelihood would require repeatedly integrating out z_i for many observations and parameter updates.

2.2.2 The Evidence Lower Bound (ELBO)

Variational inference avoids this difficulty by replacing the log marginal likelihood with a tractable lower bound (for a literature review, see Blei et al., 2017). Consider the log likelihood of the observed data $(w_{i,1:M}, y_{i,1:K})$:

$$\log p_{\theta,\psi}(w_{i,1:M}, y_{i,1:K} \mid x_i). \tag{3}$$

Let $q(z_i)$ be any distribution on $\mathcal{Z}$ that is absolutely continuous with respect to the prior. In applications, q is chosen from a tractable parametric family that is typically simpler than the true posterior (e.g., a Gaussian distribution, often with diagonal covariance, or a mean-field factorization). The evidence lower bound (ELBO) is defined as

$$\mathcal{L}_i(q) \;:=\; \mathbb{E}_{q(z_i)}\left[\sum_{m=1}^{M} \log p_{\theta_m}(w_{im} \mid z_i, x_i^c) + \sum_{k=1}^{K} \log p_{\theta_k}(y_{ik} \mid z_i, x_i^s)\right] - \mathrm{KL}\Big(q(z_i) \,\Big\|\, p_\psi(z_i \mid x_i^p)\Big). \tag{4}$$

This quantity is a lower bound on the log marginal likelihood in Eq. (3): for every choice of q,

$$\mathcal{L}_i(q) \;\le\; \log p_{\theta,\psi}(w_{i,1:M}, y_{i,1:K} \mid x_i). \tag{5}$$

The derivation of this bound is given in Appendix D.1. The gap between the left- and right-hand sides is the Kullback–Leibler divergence between $q(z_i)$ and the true posterior $p(z_i \mid w_{i,1:M}, y_{i,1:K}, x_i)$, which is always nonnegative. Hence, maximizing the ELBO pushes q toward the posterior while simultaneously fitting the modality and outcome likelihoods.

In principle, one could maximize Eq. (4) separately for each unit i by optimizing q directly (Blei et al., 2017). This would introduce unit-specific variational parameters and would be computationally expensive in large samples. Variational autoencoders instead use *amortized* inference: we learn a single function (the encoder) that maps observed data to the parameters of a variational distribution (for a literature review, see Kingma et al., 2019). In econometric terms, amortization replaces n separate nuisance optimization problems (one per observation) with a single shared inference function that can be estimated once and then evaluated cheaply for any unit. We parametrize this inference function by a neural network. Concretely, for each modality m we define a modality-specific encoder

$$q_\phi^{(m)}(z_i \mid w_{im}, x_i^p), \qquad m = 1, \dots, M,$$

where ϕ collects the encoder parameters (weights and biases). In applications, we often take $q_\phi^{(m)}$ to be a Gaussian distribution whose mean and covariance are outputs of a neural network:

$$q_\phi^{(m)}(z_i \mid w_{im}, x_i^p) = \mathcal{N}\big(\mu_\phi^{(m)}(w_{im}, x_i^p), \Sigma_\phi^{(m)}(w_{im}, x_i^p)\big).$$

Because the encoder is a neural network and the decoder is typically parameterized as a neural network, it is useful to connect the ELBO to the intuition of "encoding" and "reconstruction." The ELBO in Eq. (4) has two components. The expectation term rewards latent factors z_i that allow the decoder likelihoods to assign high probability to the observed data. The KL term penalizes variational distributions that deviate substantially from the prior $p_\psi(z_i \mid x_i^p)$. This tradeoff is the precise sense in which ELBO maximization resembles training a regularized autoencoder.

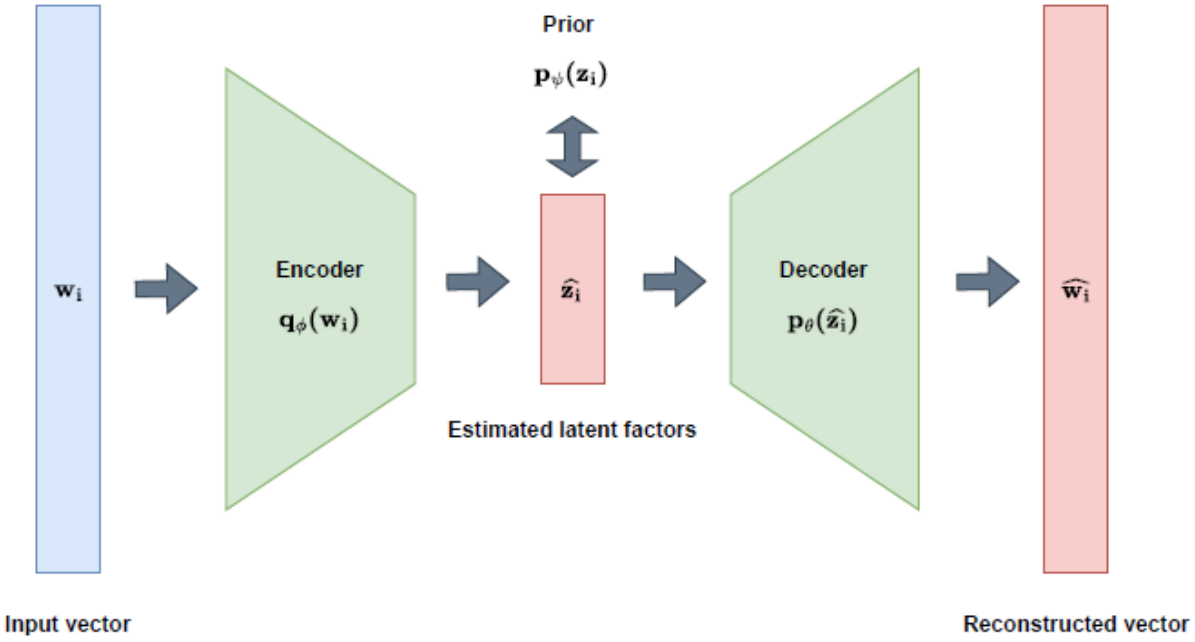
To see this in the simplest setting, consider a stylized special case with (i) no covariates, (ii) no auxiliary outcomes, and (iii) a single modality (so $M = 1$ and $K = 0$). Let $w_i := w_{i1}$ and let the model be $z_i \sim p(z)$ and $w_i \sim p_\theta(w_i \mid z_i)$. For any variational distribution $q(z_i \mid w_i)$, the ELBO reduces to

$$\mathbb{E}_{q(z_i \mid w_i)}\big[\log p_\theta(w_i \mid z_i)\big] - \mathrm{KL}\big(q(z_i \mid w_i) \,\|\, p(z_i)\big). \tag{6}$$

The first term is a probabilistic reconstruction criterion (fit of w_i given z_i), while the second term regularizes the inferred latent representation by anchoring it to the prior. In our full model, the same decomposition holds, but with multiple likelihood components (one per observed modality and one per auxiliary outcome) and with a covariate-dependent prior (see Figures 1 and 2 for a graphical overview).

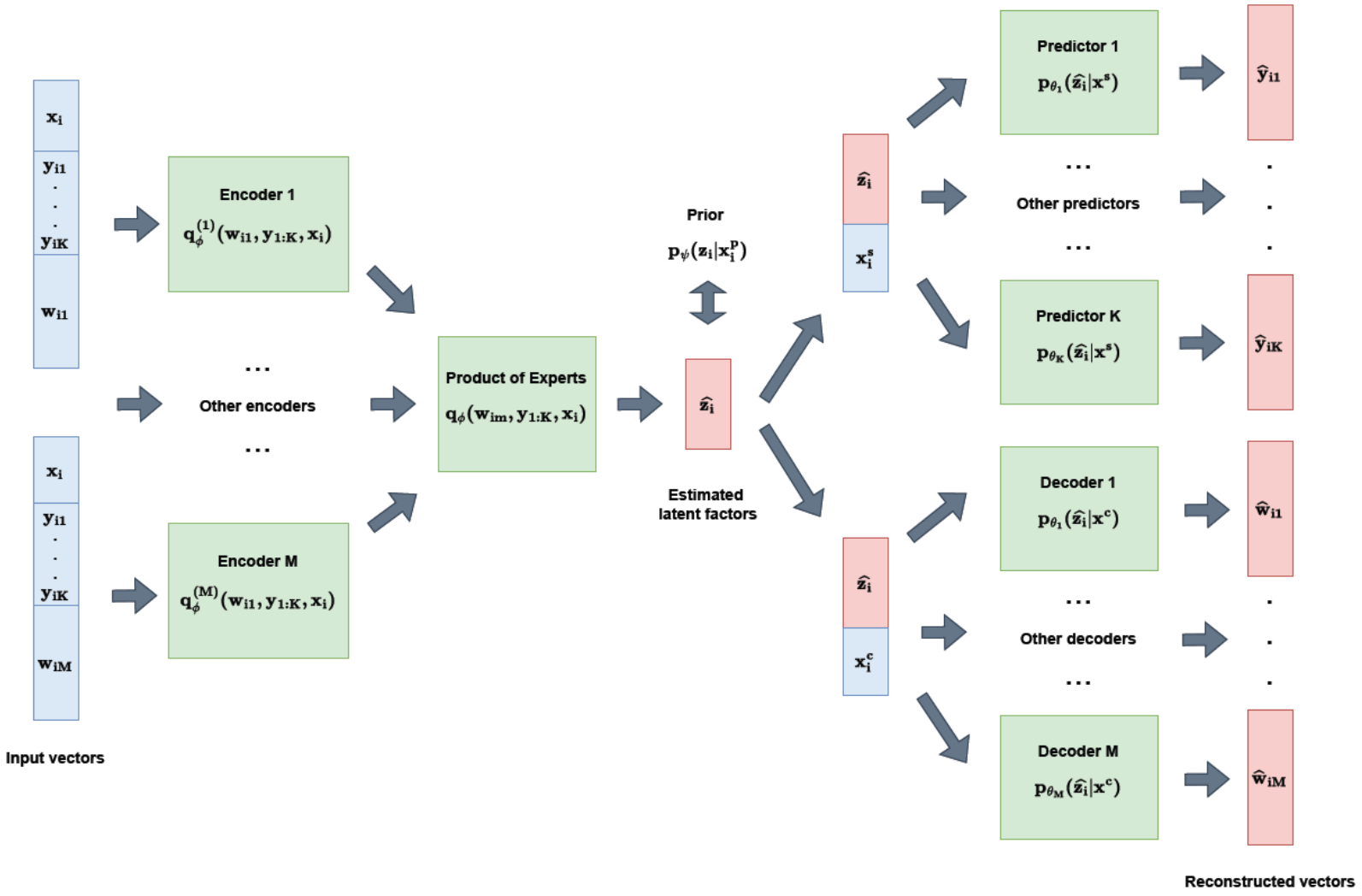
In some applications below, we replace the KL divergence with an aggregate matching penalty (e.g., MMD), yielding a Wasserstein autoencoder objective; this modification preserves the reconstruction structure while relaxing pointwise regularization of the posterior.

Figure 1: Graphical Overview of the Original Variational Autoencoder



Notes: A graphical depiction of the model architectures. Vectors colored in blue are observed. Blocks colored in green are estimated objects, typically neural networks. Vectors colored in red are the generative model's predictions.

Figure 2: Graphical Overview of our Extended Variational Autoencoder



Notes: A graphical depiction of the model architectures. Vectors colored in blue are observed. Blocks colored in green are estimated objects, typically neural networks. Vectors colored in red are the generative model's predictions.

2.2.3 Extension to missing modalities

In many applications, not all modalities are observed for all units. Let $O_i \subseteq \{1,\dots,M\}$ denote the set of modalities observed for unit i. At training time and at test time, we may further choose to condition only on a subset of the observed modalities; we denote such a conditioning set by $S \subseteq O_i$. For any $S \subseteq \{1,\dots,M\}$ we define the corresponding (subset) log marginal likelihood

$$\log p_{\theta,\psi}(w_{i,S}, y_{i,1:K} \mid x_i) \tag{7}$$

and the associated subset ELBO

$$\mathcal{L}_i(S;q) \;:=\; \mathbb{E}_{q(z_i)}\left[\sum_{m\in S} \log p_{\theta_m}(w_{im} \mid z_i, x_i^c) + \sum_{k=1}^{K} \log p_{\theta_k}(y_{ik} \mid z_i, x_i^s)\right] - \mathrm{KL}\Big(q(z_i) \,\Big\|\, p_\psi(z_i \mid x_i^p)\Big). \tag{8}$$

By construction, for every q this lower bound satisfies

$$\mathcal{L}_i(S;q) \;\le\; \log p_{\theta,\psi}(w_{i,S}, y_{i,1:K} \mid x_i). \tag{9}$$

Our model is designed for settings where units have different observed modality sets O_i and where we want inference to work even when only a subset of modalities is available. To achieve this, we train by randomly subsampling modalities. For each unit i we assume we can sample a subset $S_i \subseteq O_i$ from a user-chosen distribution $r(\cdot \mid O_i)$. In a training step we compute an ELBO that includes likelihood contributions only for the modalities in S_i (and for all auxiliary outcomes, if outcomes are always observed). Concretely, we evaluate

$$\mathcal{L}_i\big(S_i;\theta,\phi,\psi\big) := \mathcal{L}_i\Big(S_i;\, q_\phi(\cdot \mid d_{i,S_i}, y_{i,1:K}, x_i)\Big), \tag{10}$$

where $\mathcal{L}_i(S;\cdot)$ is defined in Eq. (8) and $d_{i,S_i} := \{w_{im} : m \in S_i\}$. By Eq. (9), for each realized subset S_i we have the pointwise bound

$$\mathcal{L}_i\big(S_i;\theta,\phi,\psi\big) \;\le\; \log p_{\theta,\psi}(w_{i,S_i}, y_{i,1:K} \mid x_i). \tag{11}$$

Because S_i is random during training, our training objective averages these bounds over the subsampling distribution:

$$\mathcal{J}(\theta,\phi,\psi) := \sum_{i=1}^{n} \mathbb{E}_{S_i\sim r(\cdot\mid O_i)}\Big[\mathcal{L}_i\big(S_i;\theta,\phi,\psi\big)\Big]. \tag{12}$$

Taking expectations in Eq. (11) yields

$$\mathbb{E}_{S_i\sim r(\cdot\mid O_i)}\Big[\mathcal{L}_i(S_i;\theta,\phi,\psi)\Big] \;\le\; \mathbb{E}_{S_i\sim r(\cdot\mid O_i)}\Big[\log p_{\theta,\psi}(w_{i,S_i}, y_{i,1:K} \mid x_i)\Big]. \tag{13}$$

Therefore, maximizing $\mathcal{J}(\theta,\phi,\psi)$ maximizes a lower bound on the expected log likelihood induced by the training-time random missingness pattern.

2.3 Statistical guarantees

Given that we are optimizing a lower bound on the marginal likelihood, a natural question arises: when is this bound tight so that inference is exact?

2.3.1 Inference gap: definition and decomposition

Fix a unit i and a conditioning subset $S \subseteq \{1, \dots, M\}$. Let

$$d_{i,S} := (w_{i,S}, y_{i,1:K})$$

denote the observed modalities in S together with the (continuous) auxiliary outcomes, and write $p_\vartheta(z_i \mid d_{i,S}, x_i)$ for the exact posterior under generative parameters $\vartheta := (\theta, \psi)$. For any candidate distribution $q(z_i)$ that is absolutely continuous with respect to the posterior, the standard ELBO identity gives

$$\log p_\vartheta(d_{i,S} \mid x_i) \;=\; \mathcal{L}_i(S; q) \;+ \mathrm{KL}\Big(q(z_i) \,\Big\|\, p_\vartheta(z_i \mid d_{i,S}, x_i)\Big). \tag{14}$$

We therefore define the (subset-specific) inference gap as

$$\mathrm{Gap}_i(S; q) := \log p_\vartheta(d_{i,S} \mid x_i) - \mathcal{L}_i(S; q) = \mathrm{KL}\Big(q(z_i) \,\Big\|\, p_\vartheta(z_i \mid d_{i,S}, x_i)\Big) \;\geq 0.$$

The bound is tight if and only if $q(z_i)$ coincides with the true posterior almost surely.

In amortized variational inference, q is not optimized separately for each observation. Instead one restricts attention to a tractable variational family $\mathcal{Q}$ and uses a shared encoder $q_\phi(\cdot \mid d_{i,S}, x_i) \in \mathcal{Q}$. The resulting approximation error decomposes into a *variational gap* and an *amortization gap* (Cremer et al., 2018):

$$\underbrace{\log p_\vartheta(d_{i,S} \mid x_i) - \mathcal{L}_i(S; q_\phi)}_{\text{total gap}} \;=\; \underbrace{\log p_\vartheta(d_{i,S} \mid x_i) - \mathcal{L}_i^\star(S)}_{\text{variational gap}} + \underbrace{\mathcal{L}_i^\star(S) - \mathcal{L}_i(S; q_\phi)}_{\text{amortization gap}}, \tag{15}$$

where $\mathcal{L}_i^\star(S) := \sup_{q \in \mathcal{Q}} \mathcal{L}_i(S; q)$ is the best achievable ELBO within $\mathcal{Q}$.

The *variational gap* measures the cost of restricting inference to the family $\mathcal{Q}$. By Equation (14), this gap equals the smallest Kullback–Leibler divergence from any member of $\mathcal{Q}$ to the true posterior. For this gap to vanish, two conditions must be met. First, the true posterior must lie in (or be well-approximated by) the variational family: if $\mathcal{Q}$ is not expressive enough, no amount of optimization will close the gap. Second, we must be able to *find* the best approximation within $\mathcal{Q}$.

The *amortization gap* measures the cost of using a shared encoder rather than optimizing the variational parameters separately for each observation. For this gap to vanish, two conditions must again be met. First, an *ideal inference function* must exist: there must be a mapping from observed data $(d_{i,S}, x_i)$ to the variational parameters that achieve the optimum $\mathcal{L}_i^\star(S)$. Second, we must be able to *find* (or approximate) such a mapping with the chosen encoder architecture.

When both gaps are negligible, the ELBO is tight, and inference is exact. Proposition 1 below formalizes the conditions required to bridge the inference gap.

2.3.2 When does the inference gap vanish?

Proposition 1 below gives a generic set of sufficient conditions under which the total gap can be made arbitrarily small (uniformly over a compact data domain) by choosing a sufficiently expressive variational family and encoder class.

Proposition 1 (Tightness of the ELBO). *Fix $\vartheta = (\theta, \psi)$ and a subset $S \subseteq \{1, \dots, M\}$. Suppose:*

(i) (Compact domain) *The data $(d_{i,S}, x_i)$ lie in a compact set $\mathcal{D}_S \times \mathcal{X}$, and the variational parameter space Λ is a subset of $\mathbb{R}^p$ for some $p \geq 1$.*

(ii) (Expressive variational family) *For every $(d, x) \in \mathcal{D}_S \times \mathcal{X}$,*

$$\inf_{q \in \mathcal{Q}} \mathrm{KL}\Big(q \,\Big\|\, p_\vartheta(\cdot \mid d, x)\Big) = 0.$$

(iii) (Continuous near-optimal selector in $\mathcal{Q}$) *Write $\mathcal{Q} = \{q_\lambda : \lambda \in \Lambda\}$ and define the best-in-family (per-datapoint) KL gap*

$$\Delta_{\mathcal{Q}}(d, x) := \inf_{\lambda \in \Lambda} \mathrm{KL}\Big(q_\lambda \,\Big\|\, p_\vartheta(\cdot \mid d, x)\Big).$$

For every $\delta > 0$, there exists a continuous function $\lambda^\delta : \mathcal{D}_S \times \mathcal{X} \to \Lambda$ such that for all $(d, x) \in \mathcal{D}_S \times \mathcal{X}$,

$$\mathrm{KL}\Big(q_{\lambda^\delta(d,x)} \,\Big\|\, p_\vartheta(\cdot \mid d, x)\Big) \leq \Delta_{\mathcal{Q}}(d, x) + \delta.$$

(iv) (Regularity of KL divergence) *The map*

$$(\lambda, d, x) \mapsto \mathrm{KL}\Big(q_\lambda \,\Big\|\, p_\vartheta(\cdot \mid d, x)\Big)$$

is continuous on $\Lambda \times \mathcal{D}_S \times \mathcal{X}$.

(v) (Encoder can approximate the selector) *For every $\delta > 0$, let $\lambda^\delta : \mathcal{D}_S \times \mathcal{X} \to \Lambda$ be the continuous near-optimal selector whose existence is guaranteed by condition (iii). Then for every $\eta > 0$ there exists an encoder parameterization ϕ such that*

$$\sup_{(d,x) \in \mathcal{D}_S \times \mathcal{X}} \big\| g_\phi(d, x) - \lambda^\delta(d, x) \big\| < \eta.$$

Then for every $\varepsilon > 0$ there exists an encoder parameterization ϕ such that

$$0 \;\leq\; \log p_\vartheta(d_{i,S} \mid x_i) - \mathcal{L}_i\big(S; q_\phi(\cdot \mid d_{i,S}, x_i)\big) \;\leq\; \varepsilon \quad \text{for all } (d_{i,S}, x_i) \in \mathcal{D}_S \times \mathcal{X}.$$

The proof is given in Appendix E.1. In plain words, Proposition 1 says that the ELBO can be made arbitrarily tight if (ii) the variational family is rich enough to approximate the true posterior at each data point, (iii) near-optimal variational parameters can be selected as a continuous function of the data, and (v) the encoder class can uniformly approximate this selector.

Remark 1. *The five conditions serve distinct roles:*

- *Condition (i) provides the topological structure needed for uniform approximation arguments.*
- *Condition (ii) ensures the* variational gap *can be closed: the family $\mathcal{Q}$ is rich enough to approximate the true posterior at each data point.*
- *Condition (iii) ensures the* amortization gap *can be closed in principle: optimal (or near-optimal) variational parameters can be selected as a continuous function of the data, which is necessary for a shared encoder to track them.*
- *Condition (iv) is a regularity condition ensuring that small perturbations in the variational parameters lead to small changes in the KL divergence, uniformly over the data domain.*
- *Condition (v) links the encoder architecture to the selector from condition (iii): it requires that the encoder class can uniformly approximate the particular continuous near-optimal selector λ^δ.*

Remark 2. *In our setting, an ideal inference function exists (in the sense of a mapping from $(d_{i,S}, x_i)$ to variational parameters that achieve the per-datapoint optimum within $\mathcal{Q}$). In particular, we show in Appendix D.2 that our model is a simple hierarchical model, and under the regularity assumptions in Margossian and Blei (2024) this implies condition (iii) in Proposition 1.*

Remark 3. *Condition (iv) holds under mild assumptions on the variational family and the model. For example, if $\mathcal{Q}$ is a Gaussian family with parameters $\lambda = (\mu, \Sigma)$ representing the mean and covariance, and if the posterior $p_\vartheta(\cdot \mid d, x)$ has a continuous density with respect to (d, x) and finite second moments uniformly over the compact domain, then standard continuity results for KL divergence between Gaussians (or between a Gaussian and a regular density) imply condition (iv).*

Remark 4. *Proposition 1 is an inference result: it characterizes when a shared encoder can approximate the posterior for a fixed generative model. Establishing frequentist guarantees for the estimator (consistency and asymptotic normality of ELBO maximizers) additionally requires identification and smoothness, and we illustrate this below in an extended conditional linear–Gaussian case.*

Remark 5. *In the extended conditional linear–Gaussian case introduced below, for any fixed ϑ and any subset S, the exact subset posterior $p_\vartheta(z_i \mid d_{i,S}, x_i)$ is Gaussian with a closed-form mean and covariance (Appendix E). Consequently, if $\mathcal{Q}$ contains full-covariance Gaussians, the variational gap is zero. If moreover the encoder class can represent the analytic posterior-parameter map $(d_{i,S}, x_i) \mapsto (m_{i,S}, V_{i,S})$ (or if one plugs in this map), then the amortization gap is also zero, and the total gap vanishes.*

2.3.3 A tractable special case: the linear–Gaussian VAE

This special case provides a setting in which the DGP is fully transparent and the exact posterior is available in closed form. It therefore serves to illustrate (i) how and when the ELBO gap can vanish (both variationally and through amortization) and (ii) how, under correct specification and standard regularity conditions, ELBO maximization inherits familiar consistency and asymptotic normality results from classical likelihood-based M-estimation. Throughout, let $z_i \in \mathbb{R}^L$, covariates $x_i = (x_i^p, x_i^c, x_i^s)$, modalities $w_{im} \in \mathbb{R}^{D_m}$, and continuous auxiliary outcomes $y_i := (y_{i1}, \ldots, y_{iK}) \in \mathbb{R}^K$.

Model. Fix finite-dimensional feature maps $f_p(x_i^p) \in \mathbb{R}^{r_p}$, $f_c(x_i^c) \in \mathbb{R}^{r_c}$, and $f_s(x_i^s) \in \mathbb{R}^{r_s}$ (e.g., identity maps or spline bases). The generative model specializes Eq. (1) to linear conditionals with Gaussian noise:

$$\begin{aligned}
&\textbf{Latent factors:} && z_i \mid x_i^p \sim \mathcal{N}\big(\mu_0(x_i^p), \Sigma_0\big), \qquad \mu_0(x_i^p) := B_0 f_p(x_i^p), \\
&\textbf{Modalities:} && w_{im} \mid (z_i, x_i^c) \sim \mathcal{N}\big(b_m(x_i^c) + A_m(x_i^c) z_i, \Sigma_{w,m}\big), \quad m = 1, \dots, M, \\
&\textbf{Auxiliary outcomes:} && y_i := (y_{i1}, \dots, y_{iK})^\top, \qquad y_i \mid (z_i, x_i^s) \sim \mathcal{N}\big(d(x_i^s) + C(x_i^s) z_i, \Sigma_y\big),
\end{aligned} \tag{16}$$

with conditional independence across modalities and outcomes given (z_i, x_i); in particular, the conditional-independence structure in Eq. (1) corresponds here to taking Σ_y diagonal. Here $b_m(x_i^c)$ and $A_m(x_i^c)$ are affine functions of $f_c(x_i^c)$, and $d(x_i^s)$ and $C(x_i^s)$ are affine functions of $f_s(x_i^s)$ (full parameterization in Appendix E). Let $\vartheta = (\theta, \psi)$ collect the finite-dimensional coefficients and covariance parameters.

Implications for variational inference and likelihood-based estimation. This linear–Gaussian specification is fully tractable. For any subset of observed modalities S (and with or without outcomes), the exact posterior $p_\vartheta(z_i \mid d_{i,S}, x_i)$ is Gaussian with closed-form mean and covariance (Appendix E). Consequently, a Gaussian variational family with unrestricted covariance can match the exact posterior, implying that the variational gap is zero.

The analytic posterior-parameter map $(d_{i,S}, x_i) \mapsto (m_{i,S}, V_{i,S})$ is smooth in this setting. Hence, if the encoder class can represent (or uniformly approximate) this map, amortization does not constrain inference and the amortization gap is also zero. Moreover, Bayes' rule and the conditional-independence structure of the model imply that the subset posterior aggregates information across modalities multiplicatively, with the prior factor entering exactly once. In the Gaussian case, this multiplicative aggregation implies a simple additive decomposition in terms of precision matrices. The fusion rule we introduce below in Eq. (19) is designed to mirror this multiplicative structure and is therefore exact in the present linear–Gaussian case when each unimodal encoder coincides with the corresponding unimodal posterior.

Finally, because the ELBO is tight when q coincides with the exact posterior, profiling out the variational distribution yields an objective in ϑ that shares maximizers with the marginal likelihood in this linear factor model, up to the usual orthogonal-rotation invariance. Treating the (simulated) ELBO as an M-estimation criterion then yields consistency and a $\sqrt{n}$ central limit theorem for the joint decoder–encoder maximizer, once the rotation normalization is taken into account.

Assumptions and main results. We state the formal regularity and identification conditions below; the posterior and fusion algebra (and all proofs) are in Appendix E and Appendix F.

Assumption 1 (Regularity and identification for the extended linear–Gaussian case)**.** *Let* $\eta := (\vartheta, \phi)$ *with parameter space* $\mathcal{H} = \mathcal{V} \times \Phi$. *Assume:*

(i) (Compactness / bounded eigenvalues) *$\mathcal{H}$ is compact (or η_0 is an interior point), and the eigenvalues of Σ_0, each $\Sigma_{w,m}$, and Σ_y are uniformly bounded away from 0 and ∞ on $\mathcal{H}$.*

(ii) (Sampling and covariates) *The augmented data $U_i := (d_{i,S}, x_i, \varepsilon_i)$ are i.i.d., where $\varepsilon_i \sim \mathcal{N}(0, I_L)$ is independent of $(d_{i,S}, x_i)$; covariates have bounded support (or sufficient moments) and the feature maps f_p, f_c, f_s are bounded.*

(iii) (Correct specification) *There exists $\eta_0 = (\vartheta_0, \phi_0)$ such that the model generates the distribution of $(d_{i,S}, x_i)$ and $q_{\phi_0}(\cdot \mid d_{i,S}, x_i) = p_{\vartheta_0}(\cdot \mid d_{i,S}, x_i)$ almost surely.*

(iv) (Identification up to rotations) *The generative parameters are identified up to orthogonal rotations of the latent space: for $R \in O(L)$, rotating z_i and correspondingly right-multiplying the loading matrices (and rotating the prior-mean coefficients) leaves the marginal distribution of $(d_{i,S} \mid x_i)$ unchanged. Convergence of ϑ is measured in the corresponding quotient metric (or under any imposed normalization).*

(v) (Posterior / encoder realizability) *For each ϑ in a neighborhood of ϑ_0 there exists $\phi^*(\vartheta)$ such that $q_{\phi^*(\vartheta)}(\cdot \mid d_{i,S}, x_i) = p_\vartheta(\cdot \mid d_{i,S}, x_i)$ almost surely.*

(vi) (Smoothness for CLT) *In a neighborhood of an identified representative η_0, the integrand $m(\eta; U)$ defined below is twice continuously differentiable in η almost surely, with $\mathbb{E}\,\|\nabla_\eta m(\eta_0; U)\|^2 < \infty$ and $A := -\mathbb{E}[\nabla^2_{\eta\eta} m(\eta_0; U)]$ nonsingular on identified coordinates.*

ELBO criterion as an M-estimator (one reparameterized draw). We use a Gaussian encoder $q_\phi(z_i \mid d_{i,S}, x_i)$ and the reparameterization trick, i.e., z_i is written as a deterministic function of $(d_{i,S}, x_i, \varepsilon_i)$ with $\varepsilon_i \sim \mathcal{N}(0, I_L)$. This yields a per-observation criterion contribution $m(\vartheta, \phi; d_{i,S}, x_i, \varepsilon_i)$ whose expectation over ε_i equals the exact ELBO. We maximize the corresponding one-draw simulated objective

$$\widehat{\mathcal{J}}_n(\vartheta, \phi) := \frac{1}{n} \sum_{i=1}^{n} m\big(\vartheta, \phi; d_{i,S}, x_i, \varepsilon_i\big). \tag{17}$$

The population objective $\mathcal{J}(\vartheta, \phi) := \mathbb{E}[m(\vartheta, \phi; d_S, x, \varepsilon)]$ equals the exact ELBO (expectation over ε). Keeping one draw per observation yields a simulated M-estimator; additional draws reduce simulation variance but do not change the $\sqrt{n}$ logic under standard conditions.

Theorem 1 (Consistency of the ELBO maximizer in the extended linear–Gaussian case)**.** *Suppose Assumption 1 holds and let $(\widehat{\vartheta}_n, \widehat{\phi}_n)$ be any maximizer of the objective in Eq. (17) over $\mathcal{H}$. Define the population maximizer set*

$$\mathcal{M} := \{(\vartheta, \phi) : \vartheta \in [\vartheta_0],\ q_\phi(\cdot \mid d_S, x) = p_\vartheta(\cdot \mid d_S, x)\ a.s.\},$$

where $[\vartheta_0]$ denotes the rotation-equivalence class. Then

$$\operatorname{dist}\big((\widehat{\vartheta}_n, \widehat{\phi}_n), \mathcal{M}\big) \xrightarrow{p} 0.$$

In particular, $\widehat{\vartheta}_n \to \vartheta_0$ in the quotient metric (or under the imposed normalization).

Theorem 2 ($\sqrt{n}$ rate and asymptotic normality)**.** *Under Assumption 1(vi), for an identified representative $\eta_0 = (\vartheta_0, \phi_0) \in \mathcal{M}$,*

$$\sqrt{n}(\widehat{\eta}_n - \eta_0) \Rightarrow \mathcal{N}(0, A^{-1}BA^{-1}), \qquad A := -\mathbb{E}\left[\nabla^2_{\eta\eta} m(\eta_0; U)\right], \;\; B := \mathrm{Var}\big(\nabla_\eta m(\eta_0; U)\big),$$

where $\widehat{\eta}_n := (\widehat{\vartheta}_n, \widehat{\phi}_n)$ and $U = (d_S, x, \varepsilon)$. Inference for the generative parameters ϑ uses the corresponding sub-block of $A^{-1}BA^{-1}$ (or the delta method if profiling is used).

Proofs (including the posterior algebra and Gaussian fusion exactness) are given in Appendix E and Appendix F.

2.4 Practical modeling choices

We discuss below the main modeling choices researchers face.

2.4.1 Encoders

The encoders specify the variational distribution used for approximate inference, such as $q_\phi^{(m)}(z_i \mid w_{im}, x_i^p)$ in the unimodal case and the fused distribution $q_\phi(z_i \mid d_{i,S_i}, y_{i,1:K}, x_i)$ in the multimodal case. In contrast to the true posterior, which typically has no closed form in nonlinear models, the variational distribution is chosen to be tractable and differentiable. A common choice is a Gaussian family with diagonal covariance, because it admits efficient sampling and straightforward reparameterization. More expressive choices are possible (e.g., full-covariance Gaussians, mixtures, or normalizing flows), at the cost of additional computation. In all cases, the encoder network should be understood as an estimated mapping from observables to the parameters of q_ϕ (e.g., means and variances), not as a structural component of the DGP.

From the perspective of Section 2.3, the encoder class and its output distribution jointly define the variational family $\mathcal{Q}$. Thus, choosing a richer variational family is not only a "modeling" preference; it is precisely what makes the lower bound potentially tight by reducing (or eliminating) the variational gap.

In multimodal settings, an additional modeling choice is how to construct a variational distribution that is well-defined for every subset $S_i \subseteq O_i$ of observed modalities. Define $d_{i,S_i} := \{w_{im} : m \in S_i\}$. By Bayes' rule and the conditional independence structure of the DGP,

$$p_{\theta,\psi}(z_i \mid d_{i,S_i}, y_{i,1:K}, x_i) \;\propto\; p_\psi(z_i \mid x_i^p) \left[\prod_{m \in S_i} p_{\theta_m}(w_{im} \mid z_i, x_i^c)\right] \left[\prod_{k=1}^{K} p_{\theta_k}(y_{ik} \mid z_i, x_i^s)\right]. \tag{18}$$

Equation (18) makes clear that the true posterior combines information across modalities multiplicatively (a prior factor times a product of modality likelihood factors, and outcome likelihood factors). This observation motivates our use of a product-of-experts variational family: in the idealized case where each unimodal encoder matches the corresponding unimodal posterior, the

product exactly reconstructs the subset posterior in Eq. (18). In contrast, a mixture-of-experts fusion rule would combine unimodal posteriors additively (via a convex combination). Such a family generally cannot coincide with the multiplicative posterior implied by the generative model, except in special degenerate cases, and would therefore typically entail an irreducible variational gap even under perfect optimization.

This design choice is closely related to the multimodal VAE literature. Our approach most closely resembles Wu and Goodman (2018), who combine modality-specific approximate posteriors through a product. In contrast, Suzuki et al. (2016) concatenate modality features into a single encoder, while Shi et al. (2019) use a mixture-of-experts fusion rule.

A natural strategy is to train unimodal encoders $q_\phi^{(m)}(z_i \mid w_{im}, \cdot)$ and then fuse them when multiple modalities are present. This requires a correction: each unimodal posterior contains a copy of the prior factor, so a naive product duplicates the prior and yields an overconfident approximation. We therefore fuse modality-specific encoders using a product-of-experts rule:

$$q_\phi(z_i \mid d_{i,S_i}, y_{i,1:K}, x_i) \propto \frac{\prod_{m \in S_i} q_\phi^{(m)}(z_i \mid w_{im}, y_{i,1:K}, x_i)}{p_\psi(z_i \mid x_i^p)^{|S_i|-1}}. \tag{19}$$

The denominator removes the redundant prior factors and ensures that the prior enters exactly once, mirroring Eq. (18). When the unimodal encoders are Gaussian, the PoE yields an explicit fusion rule in terms of precisions. Suppose

$$p_\psi(z_i \mid x_i^p) = \mathcal{N}(\mu_{0i}, \Lambda_{0i}^{-1}),$$

and each modality encoder has the form

$$q_\phi^{(m)}(z_i \mid w_{im}, \cdot) = \mathcal{N}\big(\mu_{im},\ (\Lambda_{0i} + \Delta_{im})^{-1}\big),$$

where $\Delta_{im} \succeq 0$ can be interpreted as the precision (information) contributed by modality m beyond the prior. Then the fused encoder for subset S_i is Gaussian with precision

$$\Lambda_{S_i} = \Lambda_{0i} + \sum_{m \in S_i} \Delta_{im},$$

and mean obtained by aggregating natural parameters. This makes clear that each modality contributes additively to information about z_i, a familiar structure in linear-Gaussian latent factor models.

2.4.2 Decoders

The generative model specifies conditional likelihoods for each modality and each auxiliary outcome:

$$p_{\theta_m}(w_{im} \mid z_i, x_i^c), \quad m = 1, \ldots, M, \qquad \text{and} \qquad p_{\theta_k}(y_{ik} \mid z_i, x_i^s), \quad k = 1, \ldots, K.$$

Here $\theta = \{\theta_m, \theta_k\}$ denotes the parameters of the decoder networks. In practice, θ_m and θ_k define the likelihood family: Gaussian for continuous data, Poisson for counts, and categorical or multinomial for text and discrete outcomes. In our implementation, both encoders and decoders are parametrized as feed-forward neural networks, but this is not a requirement. For instance, in the case of images as data, researchers may want to specify convolutional neural networks (Tolstikhin et al., 2018) or diffusion models (Ho et al., 2020). For text as data, natural choices are fully connected feed-forward neural networks for small datasets (Srivastava and Sutton, 2022; Nan et al., 2019), and transformers to model text sequences for larger datasets (Vaswani et al., 2017). Of course, the decoder can also be a fully specified structural economic model that models unobserved heterogeneity as a continuous vector of latent variables.

2.4.3 Prior

The prior is parameterized as $p_\psi(z_i \mid x_i^p)$. Its parametrization depends on the researcher's goals. To recover the standard ideal point model, one can posit a standard normal. For a typical topic model, one can posit a standard logistic normal. To increase the prior's expressivity, it can also be conditioned on covariates, such that it is a normal or logistic normal with mean $\mu_\psi(x_i^p)$ and covariance $\Sigma_\psi(x_i^p)$ (Roberts et al., 2013). It can also be modeled with more flexible parametric models—such as Gaussian mixtures or normalizing flows (e.g., Tomczak and Welling, 2018; Gatopoulos et al., 2020). Learning a structured prior may be helpful if one wishes to generate new conditional samples from the decoders.

2.5 Estimation algorithm

2.5.1 The algorithm

Table 1 summarizes the estimation procedure. We estimate parameters (θ, ϕ, ψ) by stochastic gradient ascent on the objective in Eq. (12). The ELBO involves expectations with respect to q_ϕ, which we evaluate by Monte Carlo sampling. To differentiate through Monte Carlo draws, we use the reparameterization trick (Kingma, 2013). For a Gaussian encoder, we can write

$$z_i = \mu_\phi + \Sigma_\phi^{1/2} \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

so that randomness is carried by ε rather than by ϕ. This allows gradients of the ELBO to be computed by backpropagation. Each iteration draws a minibatch of units, samples modality subsets, forms the fused encoder using Eq. (19), draws Monte Carlo samples of z_i using the reparameterization trick, and updates parameters using an adaptive optimizer such as ADAM (Kingma and Ba, 2015). Subsampling forces each modality-specific encoder to solve the inference problem even when other modalities are absent. Without subsampling, a common failure mode is that the training process concentrates inference capacity on the most informative modality and leaves other encoders weak, which then produces unreliable out-of-sample inference when only a "non-dominant" modality is observed.

Table 1: Estimation algorithm

Input: Dataset $\{(w_{i,1:M}, y_{i,1:K}, x_i)\}_{i=1}^N$

Initialize: Encoders $\{\phi_m\}_{m=1}^M$; decoders $\{\theta_m\}_{m=1}^M$; predictors $\{\theta_k\}_{k=1}^K$; prior ψ.

for epoch $= 1, \ldots, E$ **do**

  for minibatch $\{(w_{i,1:M}, y_{i,1:K}, x_i)\}_{i=1}^n$ **do**

    Randomly sample modality subset $S_i \subseteq \{1, \ldots, M\}$ for each i.

    For each $m \in S_i$, compute $q_{\phi_m}(z_i \mid w_{im}, y_{i,1:K}, x_i)$.

    Aggregate via PoE:

$$q_\phi(z_i \mid d_i, x_i) \propto \frac{\prod_{m \in S_i} q_{\phi_m}(z_i \mid w_{im}, y_{i,1:K}, x_i)}{p_\psi(z_i \mid x_i^p)^{|S_i|-1}}.$$

    Sample $\hat{z}_i \sim q_\phi(z_i \mid d_i, x_i)$ via reparameterization.

    Evaluate $\log p_{\theta_m}(w_{im} \mid \hat{z}_i, x_i^c)$ for all m, and $\log p_{\theta_k}(y_{ik} \mid \hat{z}_i, x_i^s)$ for all k.

    Compute minibatch ELBO $\widehat{\mathcal{L}}$.

    Update $(\{\phi_m\}, \{\theta_m\}, \{\theta_k\}, \psi)$ via SGD on $\widehat{\mathcal{L}}$.

  end for

end for

This estimation framework is computationally attractive relative to both MCMC and traditional (non-amortized) variational inference. Relative to MCMC, the approach avoids generating long correlated Markov chains for each dataset and instead uses cheap Monte Carlo samples inside minibatch stochastic gradients; training therefore scales well with n through subsampling, is straightforward to parallelize on GPUs, and does not require repeated full-data posterior sweeps. Relative to classical variational inference that optimizes unit-specific variational parameters (e.g., per-document topic proportions) in an inner loop, amortization replaces iterative per-unit updates with a single forward pass through the encoder, so memory does not grow with n and out-of-sample inference reduces to evaluating a trained network. As a result, we obtain an estimator that is well-suited to large samples and nonconjugate likelihoods while retaining a likelihood-based objective.

2.5.2 Posterior collapse

A recurrent challenge in training VAEs is *posterior collapse*. Posterior collapse occurs when the approximate posterior distribution learned by the encoder converges to the prior distribution regardless of the input data. In this case, the latent variables carry no information about the observations, and the decoder generates outputs almost entirely without using the latent code. The KL divergence term in the ELBO then approaches zero, signaling that the latent representation has become redundant. This undermines the purpose of the VAE, since the model no longer learns a meaningful mapping between data and latent space. In the linear Gaussian VAE setting, it can be

shown that posterior collapse occurs when the model converges to a local minimum (Lucas et al., 2019). In fact, posterior collapse arises if and only if the latent factors are not identifiable—as is the case in the linear Gaussian VAE setting (Wang et al., 2023).

To mitigate posterior collapse, researchers have adopted various strategies. Srivastava and Sutton (2017) suggests training models with a high learning rate, a high momentum, and dropout. Bowman et al. (2016) proposes KL-annealing, a procedure that gradually increases the weight on the KL divergence in the ELBO from zero to one as training progresses. One can also hardcode that the KL divergence should always exceed a threshold to prevent it from going to zero—a technique known as "free bits" (Kingma et al., 2016). Finally, richer priors or more expressive variational families often reduce the odds of posterior collapse (Rezende and Mohamed, 2015; Kingma et al., 2016; Tomczak and Welling, 2018). An alternative to the VAE is the Wasserstein Autoencoder (WAE, Tolstikhin et al., 2018), which can be more stable in practice but comes at the expense of weaker likelihood-based guarantees (see Appendix C).

2.6 Useful extensions

We consider several natural extensions to the theoretical framework outlined above.

2.6.1 Weighted ELBO

A popular generalization of the ELBO is a weighted objective that weights the contributions of losses differently and scales the regularizer:

$$\begin{aligned}\mathcal{L}_{\beta,\gamma}(\theta,\phi,\psi;i) = \sum_{m=1}^{M} \mathbb{E}_{q_\phi}\left[\log p_{\theta_m}(w_{im} \mid z_i, x_i^c)\right] \; + \; \sum_{k=1}^{K} \gamma_k \, \mathbb{E}_{q_\phi}\left[\log p_{\theta_k}(y_{ik} \mid z_i, x_i^s)\right] \\ - \beta \, \mathrm{KL}\Big(q_\phi(z_i \mid w_{i,1:M}, y_{i,1:K}, x_i) \,\Big\|\, p_\psi(z_i \mid x_i^p)\Big),\end{aligned} \tag{20}$$

with $0 \leq \gamma_k$ and $0 < \beta$. When $\gamma_k = 1$ and $\beta = 1$ we recover the ELBO. For $\beta \neq 1$, the KL penalty is rescaled, no longer corresponding exactly to a likelihood bound, but can improve the interpretability of the latent space. $\beta > 1$ strengthens the KL penalty in favor of the prior, whereas $\beta < 1$ relaxes the regularization in favor of reconstruction accuracy (Higgins et al., 2017). Theoretically, $\beta < 1$ produces an objective that is no longer a guaranteed lower bound on the marginal likelihood.

An instructive special case is when the reconstruction term corresponds to mean squared error, i.e., a Gaussian likelihood with fixed variance. Since maximizing $\mathcal{L}_{\beta,\gamma}$ is equivalent (up to scaling by $1/\beta$) to maximizing a tempered-likelihood ELBO with likelihood factor $p(w \mid z)^{1/\beta}$, in the Gaussian case this corresponds (up to additive constants) to an ELBO for a model with a different observation-noise variance. In this sense, $\beta < 1$ can still be interpreted as a likelihood bound for a nearby Gaussian model with a tighter reconstruction term.

Empirically, $\beta < 1$ can mitigate posterior collapse and provide more interpretable latent factors (especially with restrictive variational families such as Gaussians). The weights γ_k are included to control the weight given to the prediction loss. This is useful if researchers prioritize achieving

high predictive accuracy over latent factor interpretability. Higher values will nudge the model to learn latent factors that are more predictive of the outcomes (Kingma et al., 2014).

2.6.2 Transfer and cross-modal learning

A natural extension of the standard VAE framework is to allow the encoder to condition on a different set of inputs than the variables appearing in the decoder likelihood. Fix a set of *decoded targets* $S^{\text{dec}} \subseteq \{1, \ldots, M\}$, and let u_i denote the information fed to the encoder (e.g., a different modality subset $w_{i,S^{\text{enc}}}$, or a learned / pre-trained representation such as an embedding). Consider the generative model for the decoded targets,

$$p_{\theta,\psi}(w_{i,S^{\text{dec}}} \mid x_i) = \int p_\psi(z_i \mid x_i^p) \prod_{m \in S^{\text{dec}}} p_{\theta_m}(w_{im} \mid z_i, x_i^c)\, \mathrm{d}z_i.$$

For any variational family of the form $q_\phi(z_i \mid u_i, x_i)$, the corresponding objective

$$\mathcal{L}_i^{\text{cross}}(S^{\text{dec}}; u_i) := \mathbb{E}_{q_\phi(z_i \mid u_i, x_i)} \left[\sum_{m \in S^{\text{dec}}} \log p_{\theta_m}(w_{im} \mid z_i, x_i^c) \right] - \mathrm{KL}\Big(q_\phi(z_i \mid u_i, x_i) \,\Big\|\, p_\psi(z_i \mid x_i^p) \Big) \tag{21}$$

is an evidence lower bound (ELBO) on the marginal log likelihood of the decoded targets:

$$\mathcal{L}_i^{\text{cross}}(S^{\text{dec}}; u_i) \;\le\; \log p_{\theta,\psi}(w_{i,S^{\text{dec}}} \mid x_i).$$

Intuitively, the encoder inputs u_i act as side information that helps construct a good variational approximation for z_i, while the likelihood term evaluates model fit only for the decoded targets in S^{dec}. If one additionally includes likelihood terms for the encoder-side variables, this reduces to the standard multimodal ELBO for the joint likelihood over all modeled modalities.

This framework covers two common use cases: *transfer learning* and *cross-modal prediction*. In transfer learning, u_i is a representation (often pre-trained) that is designed to be comparable across domains, such as multilingual text embeddings. For example, Bianchi et al. (2021) leverage multilingual text embeddings for 'zero-shot" topic modeling: they train a VAE on English documents using multilingual embeddings as encoder input, then predict topic shares for documents in other languages (e.g., German) at test time without retraining. The model generalizes across languages because the embeddings encode cross-lingual semantic similarity. In our notation, the decoder likelihood models an English document representation $w_{i,S^{\text{dec}}}$, while the encoder conditions on $u_i = e_i$ computed from any language. Equation (21) shows that training remains likelihood-based: the objective lower-bounds the marginal likelihood of the decoder targets. At test time, one can infer z_i for out-of-domain inputs via $q_\phi(z_i \mid e_i, x_i)$; however, to generate samples in the new domain, one must also specify and train a decoder for that domain.

In cross-modal prediction, u_i consists of one or more modalities, and the decoded targets are different modalities. For instance, in legislative studies, one may encode the text of speeches and decode roll-call votes (a "speech-to-vote" model): $u_i = w_{i,S^{\text{enc}}}$ is speech text (or its embedding),

and S^{dec} indexes vote outcomes. The objective in Equation (21) is then an ELBO on the marginal likelihood of votes under a latent-factor model, with the latent representation z_i constrained to be predictable from the encoded text and useful for explaining voting behavior. Similar setups arise when encoding product descriptions to decode consumer reviews, or encoding news text to decode market outcomes.

When the modalities in u_i are truly generated by the same latent factor z_i and are conditionally independent given z_i (as in our DGP), this cross-modal setup is conceptually coherent: regardless of which modality subset is observed, inference targets the same latent object, and (in expectation) different modality-specific encoders should agree on the posterior information about z_i.

3 Simulations and Common Datasets

We conduct a series of Monte Carlo exercises to assess the performance of our estimator in finite samples. We find that the GTM correctly infers document-topic shares in monolingual and multilingual settings, correctly learns the prior conditional on document metadata (i.e., the effects of covariates on topics as outcomes), and correctly captures the effects of topics as regressors. We also report the classification accuracy of our model on standard evaluation datasets from the computer science literature. In all our simulations, the dropout rate for the GTM is set to 0. For real datasets, we keep the default dropout rate of 20%.

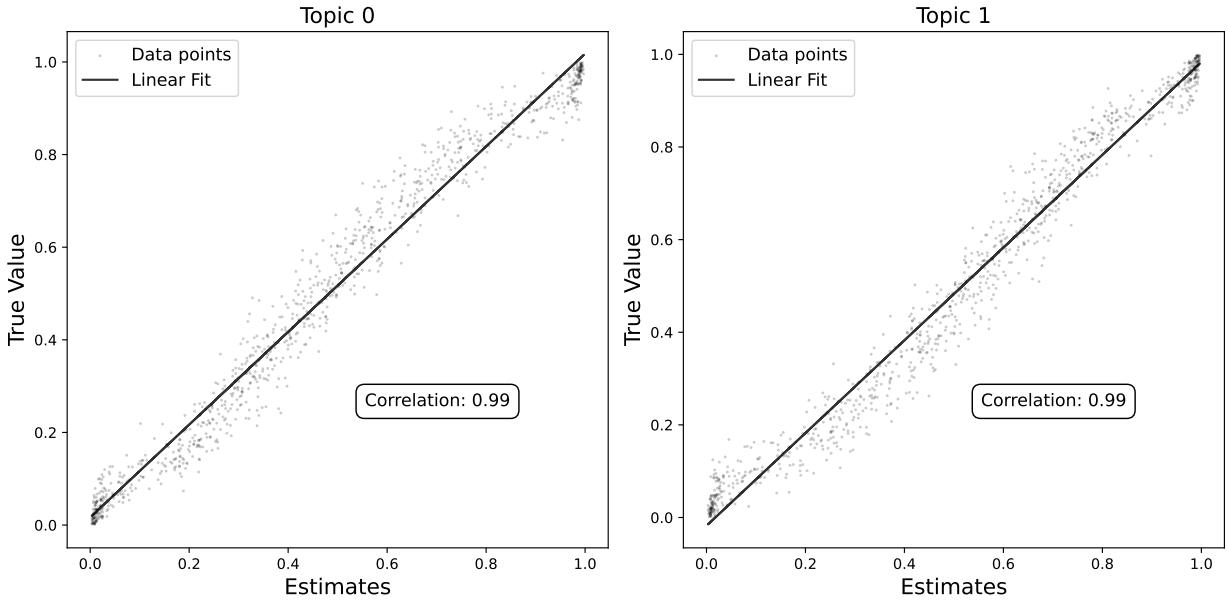
3.1 Monolingual Setting

Simulation. To assess the GTM's ability to estimate document-topic shares in a monolingual setting, we simulate a dataset of 1000 documents without document metadata. Each document comprises 100 words drawn from a vocabulary of 500 words. The prior document topic shares follow a Dirichlet distribution with a concentration parameter of 0.1, with two topics.

Figure 3 plots the estimated topic shares against the true values. The GTM performs very well in recovering both topics. The correlation between estimated document-topic shares and their true values is 0.99.

20NewsGroup Dataset. To assess the GTM's performance on a real dataset, we estimate document topic shares on the 20NewsGroup dataset, a popular collection of about 20,000 news articles organized into 20 distinct categories (e.g. politics, science, sports). 20NewsGroup is commonly used as a benchmark for evaluating algorithms on topic classification and document clustering. For our evaluation, we assess whether the most prevalent estimated topic matches with the document's true label. We find that the GTM correctly classifies 43% of documents, in line with LDA's accuracy of 42% on the same dataset. Note that some topics are very similar to one another (e.g., "electronics", "hardware", and "computer graphics") and that this explains the vast majority of misclassified articles for both models.

Figure 3: Simulation: True vs. Estimated Document-Topic Shares (Monolingual Setting)



Notes: Scatter plots of true values (vertical axis) against model estimates (horizontal axis) for topic 0 (left panel) and topic 1 (right panel), with line of best fit. Topics learned from simulated dataset of 1000 documents, each with 100 words drawn from a vocabulary of 500 words, with two topics.

3.2 Multilingual Setting

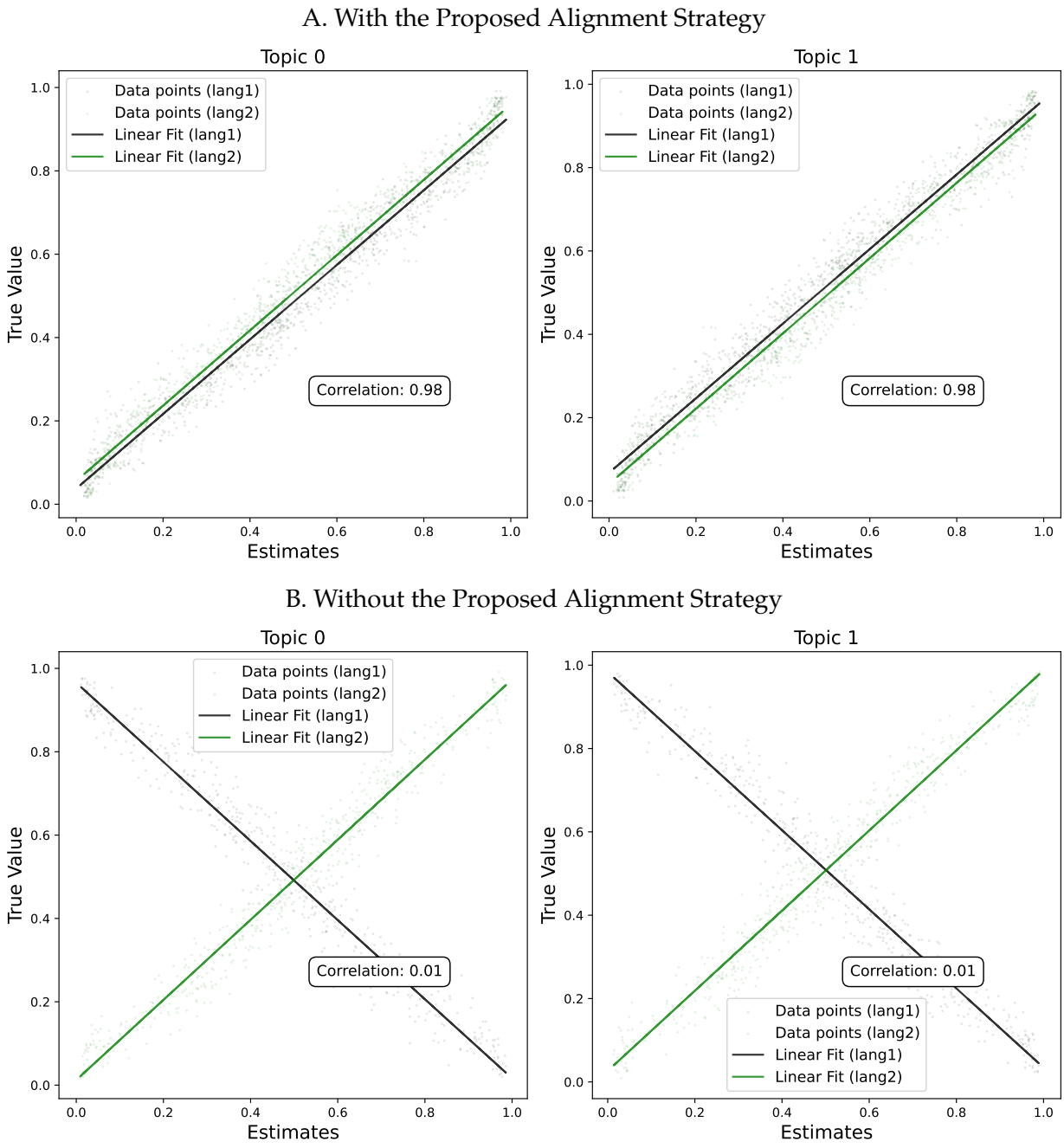
Simulation. To assess the GTM's ability to estimate document-topic shares in a multilingual setting, we adjust the simulation dataset to have two topics and two languages. As before, we simulate a dataset of 1000 documents without document metadata, where each document comprises 100 words drawn from a vocabulary of 500 words. The prior on document topic shares is a Dirichlet distribution with a concentration parameter of 0.1.

Figure 4 plots the estimated document-topic shares against their true values. Panel A corresponds to our alignment strategy, where the joint encoder learns aligned topic representations of the documents. In that case, performance is again excellent with correlations between truth and prediction of 0.98. Panel B corresponds to a naive approach that would not align topics across languages. Without the alignment strategy, the GTM estimates topics that are not aligned across languages. That is, Topic 0 in language 1 does not necessarily refer to Topic 0 in language 2.

WikiData abstracts. To assess the GTM's performance on a real dataset, we estimate document topic shares on a corpus of English WikiData abstracts. These abstracts contain summaries of Wikipedia articles on six pre-specified topics (food, religion, movies, music, politics, nature). We translate each document into Chinese using Google Translate to create an aligned multilingual corpus.

We train GTM with the encoder and the decoder both having one hidden layer with 256 units and ReLU activation. The prior on document-topic shares is a Dirichlet with concentration pa-

Figure 4: Simulation: True vs. Estimated Document-Topic Shares (Multilingual Setting)



Notes: Scatter plots of true values (vertical axis) against model estimates (horizontal axis) for topic 0 (left panel) and topic 1 (right panel), with line of best fit. Topics learned from simulated multilingual dataset of 1000 documents, each with 100 words drawn from a vocabulary of 500 words, with two topics. Panel A applies the multilingual alignment strategy, while Panel B does not.

Table 2: Classification Accuracy on the Multilingual WikiData Corpus

Input	Train Language(s)	Test Language(s)	% of Aligned Docs	Accuracy (%)
Bag of Words	English	English	-	91 (1.39)
	Chinese	Chinese	-	85 (2.15)
	English & Chinese	English & Chinese	100	85 (1.43)
			50	84 (0.85)
			10	80 (1.63)
Embeddings	English	English	0	76 (4.08)
	English	Chinese	0	72 (2.56)
	Chinese	Chinese	0	75 (2.97)
	Chinese	English	0	77 (1.54)

Notes: Multilingual classification results for a collection of 11,043 Wikipedia articles in Chinese and English from six pre-specified topics (food, religion, movies, music, politics, nature). Columns give the input format, train and test languages, number of translated documents used for alignment, and classification accuracy. We run each model with ten different starting seeds. Standard errors are reported in parentheses.

rameter 0.05. To better account for uncertainty, we run each model with ten different starting seeds.

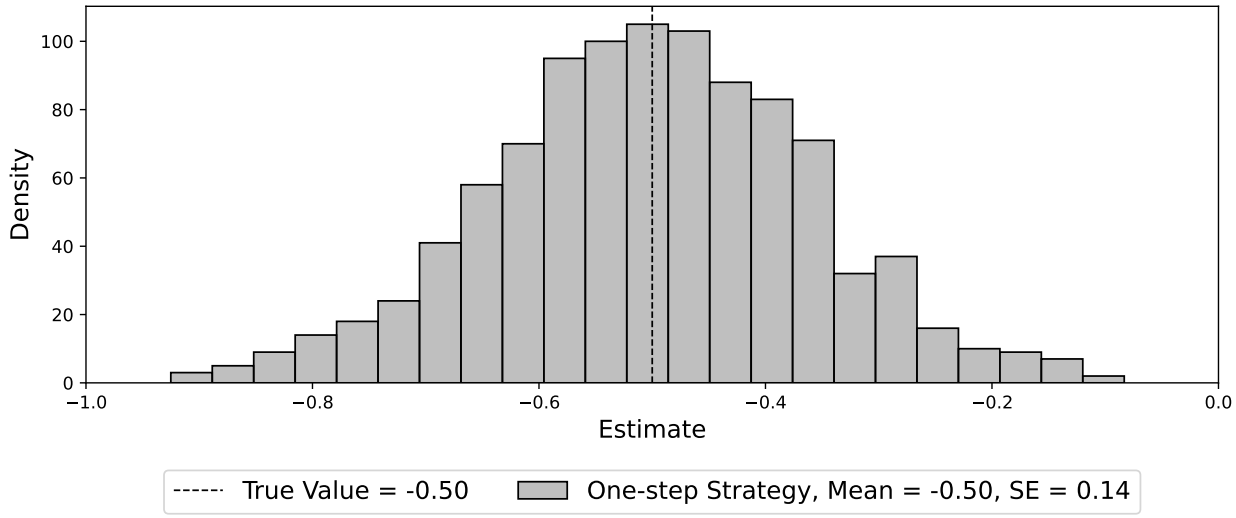
Results are presented in Table 2, where we assess the different approaches to multilingual topic modeling discussed in Section **??** above. The top panel uses the bag-of-words inputs in the original languages. First, for comparison we estimate models trained and tested in the same language. These models display 91% accuracy for English and 85% accuracy for Chinese. We then estimate models trained on both languages simultaneously with a joint encoder, but we vary the percentage of aligned documents in the datasets. These models display similar performance as their monolingual counterparts. Interestingly, the share of aligned documents barely impacts performance, suggesting a small amount of translated documents may be sufficient to align topics across languages successfully.

In the bottom panel of Table 2, we explore the zero-shot approach using multilingual document embeddings as input, where topics are learned in one language and then extrapolate to new documents in another language. To encode the documents, we use the "paraphrase-multilingual-mpnet-base-v2" model from the `sentence_transformers` Python package. Consistent with Bianchi et al. (2021), we find slightly lower accuracy for these models, compared to the bag-of-words approach using an aligned corpus. However, interestingly, the embedding-based model performs just as well across languages as it does within-language.

3.3 Topic Shares as Outcomes

Next we adapt the simulation approach to assess the ability of the GTM to capture the effect of topics as outcomes. As before, we simulate 1000 datasets of 1000 documents, where each document comprises 100 words drawn from a vocabulary of 500 words. Document topic shares are drawn from a logistic normal prior, with two topics, and with one covariate affecting document-topic

Figure 5: Simulation: Density of Estimates for the Effect of Covariates on the Prior



Notes: Histogram of estimates for the effect of a covariate that changes document-topic shares, where the true effect on the learned prior (-0.5) is a dashed line. Topics learned from simulated dataset of 1000 documents, each with 100 words drawn from a vocabulary of 500 words, with two topics.

shares.

Figure 5 shows the simulation results. We report the distribution of estimates for the effect of the covariate on topic shares from these 1000 iterations. On average, the GTM recovers the true effect of covariates on the prior.

3.4 Topic Shares as Regressors

Now we assess the ability of the GTM to capture the effect of topics as regressors. We simulate 1000 datasets of 1000 documents, where again each document comprises 100 words drawn from a vocabulary of 500 words. Document topic shares are drawn from a Dirichlet distribution with a concentration parameter 0.1, again with two topics. Document-topic shares affect one outcome.

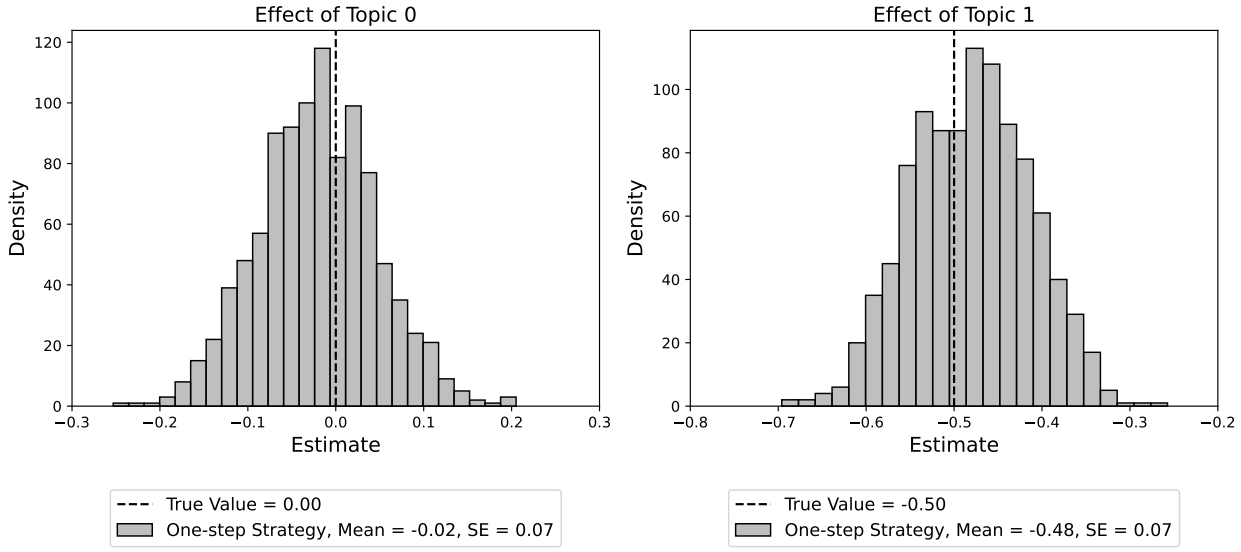
Figure 6 shows the estimation results based on the simulated data. We report the distribution of estimates from the 1000 iterations. On average, the GTM recovers the true effect of topic shares on the outcome.

3.5 Speed and Scalability

Our last set of simulations demonstrate the speed and scalability of the GTM. For these experiments, the GTM is parametrized as follows. The encoder has two hidden layers of 256 and 64 dimensions. The decoder has two hidden layers of 64 and 256 dimensions. The dropout rate to prevent overfitting is set at 20%. We stop the estimation routine when the loss fails to improve.

Figure 7 shows the time it takes to produce estimates for the GTM depending on dataset characteristics. Holding other dimensions fixed, we vary the number of documents (Panel A), the number of topics (Panel B), and the number of covariates (Panel C). The GTM's compute time

Figure 6: Simulation: Density of Estimates for the Effect of Topic Shares on the Outcome



Notes: Histogram of estimates for the effect of topic 0 (left panel) and topic 1 (right panel) on metadata outcome, where the respective true effects (0.0 for topic 0; -0.5 for topic 1) are indicated by dashed lines. Topics learned from simulated dataset of 1000 documents, each with 100 words drawn from a vocabulary of 500 words, with two topics.

barely increases along all dimensions we consider, and can be estimated on 100,000 documents in under three minutes with a GPU and under ten minutes with a CPU. This highlights that our model can easily handle very large datasets. In the application to the *U.S. Congressional Record*, for instance, we fit the model on 500,000 political speeches (see Section 4).

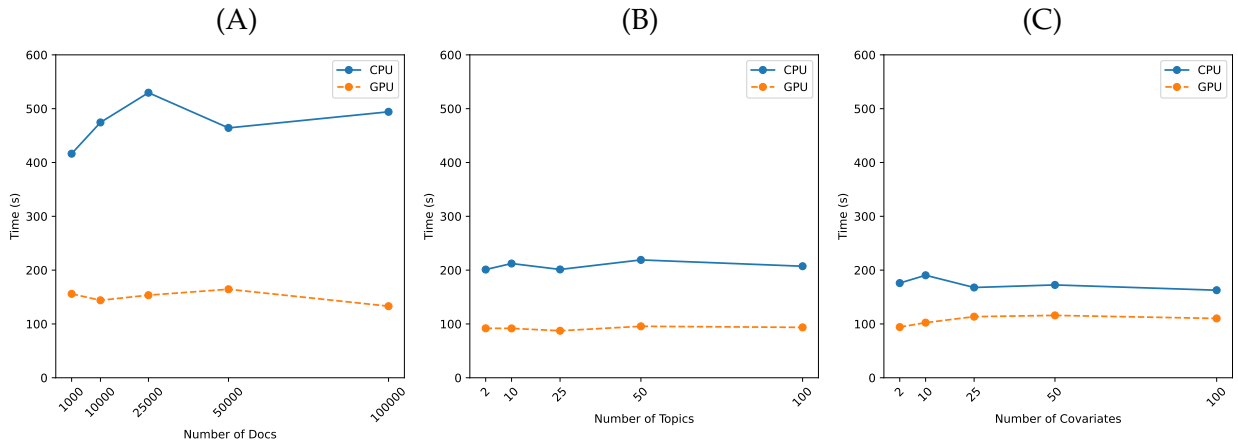
4 Application 1: Speech polarization in the US Congress

We now highlight the GTM's usefulness in an application to speech polarization on the floor of the US Congress. Understanding whether differences in speech between two groups are due to differences in the topics discussed or their respective framing is a recurring challenge in political and media economics. For instance, a well-established stylized fact in U.S. politics is that speech polarization between Republicans and Democrats has increased in recent decades. Is this increase driven by topic choice or framing? In other words, are Republicans and Democrats increasingly talking about different topics or talking about the same topics in different ways? The GTM offers a simple, disciplined, unsupervised approach to perform this decomposition exercise.

4.1 Data and methods

Data. Our application uses the *United States Congressional Record* for 1873—2016. The *Record* is an exact transcript of speeches given on the House and Senate floor. The speeches up until 1993 are from HeinOnline. As they performed optical character recognition (OCR) on scanned print volumes, the text is of good quality overall but has misrecognition errors (some letters are

Figure 7: Compute Time for Different Dataset Sizes



Notes: Plots of GTM compute time (in seconds, vertical axis) as functions of the number of documents (Panel A), number of topics (Panel B), and number of covariates (Panel C). Includes times for CPU (blue solid) and GPU (orange dashed).

sometimes mistaken for others). Since 1994, the speeches have been published in digital format by the U.S. Government Publishing Office (GPO), and the resulting texts tend to be of higher quality. We enrich the metadata associated with each speech (notably its time stamp and session) with the speaker's characteristics (in particular party affiliation).

We process the text as follows. We lowercase all words and keep proper nouns, common nouns, adjectives, and verbs. We drop a small, curated list of procedural terms. We tokenize speeches by segmenting words. We drop all speeches of less than 30 words as they are often uninformative, procedural comments. We construct unigrams, bigrams, and trigrams. Finally, we remove outlier words from the vocabulary that appear in less than 0.1% or more than 1% of speeches. The curated dataset represents approximately 3 million documents and a vocabulary of 14,503 unique words and phrases.

Estimation details. In our baseline specification, we estimate the generalized topic model for 50 topics. We specify a Logistic Normal prior for document-topic shares, no hidden layer for the encoder, and a shallow neural network with one hidden layer and no activation function for the decoder. We allow topic choice and content to depend on the same set of covariates: the time of the speech (session or biennium), the speaker's party, and party-biennium interactions.

During training, a random sample of 400,000 training documents is selected with 100,000 documents reserved for validation. To mitigate overfitting, we include a dropout rate of 20% while training. We stop updating parameters when the validation loss fails to improve.

4.2 Descriptive results

We begin with descriptive statistics on the estimated topics and embeddings. First, Table 3 lists the fifty learned topics and their most representative words and human labels. Similar to the

standard LDA model, the topics are easily interpretable. There are clear topics associated with economic and social policies—for example, labor market, taxation, pensions, education, public health, health insurance, and elections. To build some intuition for these topics, Figure 8 shows a two-dimensional mapping of the speeches represented as vectors of topic shares. Each dot represents a speech with colors assigned according to their main topic. Ultimately, the topics simply constitute soft clusters of the speeches, as can be seen from their spatial distribution in the latent factor space.

Next, we use the GTM to examine differences across Democrats and Republicans in their topic choice. Figure 9 reports the marginal effect of being a Republican on the probability of mentioning a specific topic. We can see, for example, that legal matters and foreign aid topics tend to be discussed by Democrats. Issues related to banking and finance, or to military service and veterans, tend to be discussed by Republicans. Despite these differences, the vast majority of estimated topics are equally discussed by both parties.

Now we turn to the GTM's capacity to learn associations between wording and metadata. Table 5 displays the top words associated with various points in time. These are computed by setting all topic shares and covariates to zero with the exception of the binary variable for each respective biennium, and showing the most likely words to be generated by the decoder. The first row—the words associated with the intercept term—illustrates how the inclusion of an intercept covariate in the GTM works to absorb stopwords and other generic language that is used in most speeches. Next, we have a row for a selection of years. They recapitulate distinctive events and policy priorities from the associated periods, including World War I and aftermath (1915-1921), the Great Depression and New Deal (1929-1931), World War II and aftermath (1939-1947), the Energy Crisis and Cold War (1973-1981), the War on Terror (2001-2007), and the Financial Crisis (2009-2011).

The GTM learns not just overall associations with document metadata, but also interactions between these covariates and the topics. To illustrate this, Table 6 provides examples of party-specific deviations in wording from a selection of topics. For instance, in the "Welfare and Unemployment" topic, Republicans emphasize money and pensions, while Democrats emphasize unemployment, food stamps, and poverty. In the "Elections and Voting" topic, Republicans mention terms like precinct and elector, whereas Democrats focus on civil rights and equal protection. In the "Labor Market" topic, Republicans emphasize laborers, whereas Democrats advocate for an increase in the minimum wage. In the "Crime and Justice" topic, Republicans link crime to alcohol and illegal immigrants ("alien"), whereas Democrats discuss juvenile justice and drugs. These within-topic differences in word choice capture intuitive differences in topic framing across party lines.

Finally, contrary to classical topic models, the GTM is ultimately a neural network and thus learns representations of the data in its hidden layers. For instance, with our parametrization, the decoder has a hidden layer of 256 dimensions. This layer jointly learns embeddings of documents, topics, words, and covariates. We show that the resulting embeddings space captures meaningful semantic information. Table 7 presents the closest words—as measured by cosine similarity—to

a set of common policy words in the decoder's embeddings space. The GTM has clearly learnt high-quality word embeddings in the sense that words with similar semantic meaning are located closer to one another. Going beyond these standard properties of word embeddings (e.g. Rodriguez and Spirling, 2022), the GTM also allows word representations to be directly compared to representations for documents, topics, and covariates.

4.3 Decomposition of speech polarization

An attractive property of our structural generative model of speech is that it intuitively separates topic choice from topic content. Our approach is to generate pseudo-documents based on our estimated topic model and then estimate a speech polarization index based on these newly generated documents. By generating pseudo-documents, we can compare observed speech polarization levels with hypothetical counterfactuals: (i) one in which party affiliation only influences topic choice, and (ii) one in which party affiliation only influences topic content. As sanity checks, we consider two additional counterfactuals: (iii) one in which party affiliation influences both topic choice and topic content (which should closely mirror observed speech polarization levels if the topic model is well-estimated), and (iv) one in which party affiliation influences neither topic choice nor topic content (which should translate into no speech polarization if the speech polarization index is well-estimated).

Following Gentzkow et al. (2019), we define speech polarization as *"the posterior probability that an observer with a neutral prior expects to assign to a speaker's true party after hearing the speaker utter a single phrase."* Formally, let q_{jt}^R (q_{jt}^D) denote the probability of a Republican (Democrat) pronouncing the word j in session t. The posterior belief that an observer with a neutral prior assigns to a speaker being Republican if the speaker chooses word j in session t is:

$$\rho_{jt} = \frac{q_{jt}^R}{q_{jt}^R + q_{jt}^D}.$$

We are interested in the vector of ρ_{jt}'s, $\boldsymbol{\rho_t}$, that averages this posterior belief across parties and phrases for session t. We take the estimated probability distributions of q_t^R and q_t^D from our topic model for each session. We generate 1000 speeches for each party and session under our four scenarios and compute $\boldsymbol{\rho_t}$. We do this exercise 200 times to construct confidence intervals (in practice, confidence intervals are small and therefore not shown).

The probabilities q_{jt}^D and q_{jt}^R are unobserved and need to be estimated from the data. A naive MLE estimator of these probabilities drastically overestimates speech polarization levels. This is due to the high dimensionality of the choice set (the vocabulary of words to choose from) relative to the number of words spoken in a speech (the finite number of observed realizations). A central contribution of Gentzkow et al. (2019) is to propose two estimators that mitigate this finite-sample bias. The GTM is not sensitive to this bias because it outputs the entire estimated probability distribution over words when sampling from the prior. That is, we do not generate documents with a few words drawn from the distribution of words. Instead, we take the distribution directly

to measure speech polarization. Thus, the only sampling uncertainty comes from sampling topics from the prior. Furthermore, the GTM is heavily regularized (with the dropout rate) when it learns partisan probability distributions over words, making it unlikely to capture spurious associations between party and word choice. This is reminiscent of Gentzkow et al. (2019), given that their preferred estimator penalizes party loadings.

The main result on topic decomposition is presented in Figure 10. Consistent with previous work, we find stable or even decreasing in polarization in the early years, followed by a significant increase starting in the 1990s. We find that framing (in blue) explains the entirety of the increased speech polarization (in red) in the time series. Differences in topic choice (green) have always been low, and have not changed much over the decades. This is broadly consistent with our descriptive results, which already suggested that topic choice was mostly nonpartisan.

At the bottom of graph, the time series in grey gives a placebo test, where polarization is measured from simulated speeches without the party covariates turned on. The time series clearly displays nonexistent speech polarization levels, suggesting our speech polarization estimates are not spurious or otherwise due to finite-sample bias. This placebo gives us more confidence that the main polarization measures are valid. Finally, our estimates of total polarization closely mirror the estimates of Gentzkow et al. (2019) in levels and trends. That adds some additional confidence as it replicates results from another validated approach to measuring speech polarization.

Table 3: List of Estimated Topics from the *U.S. Congressional Record*

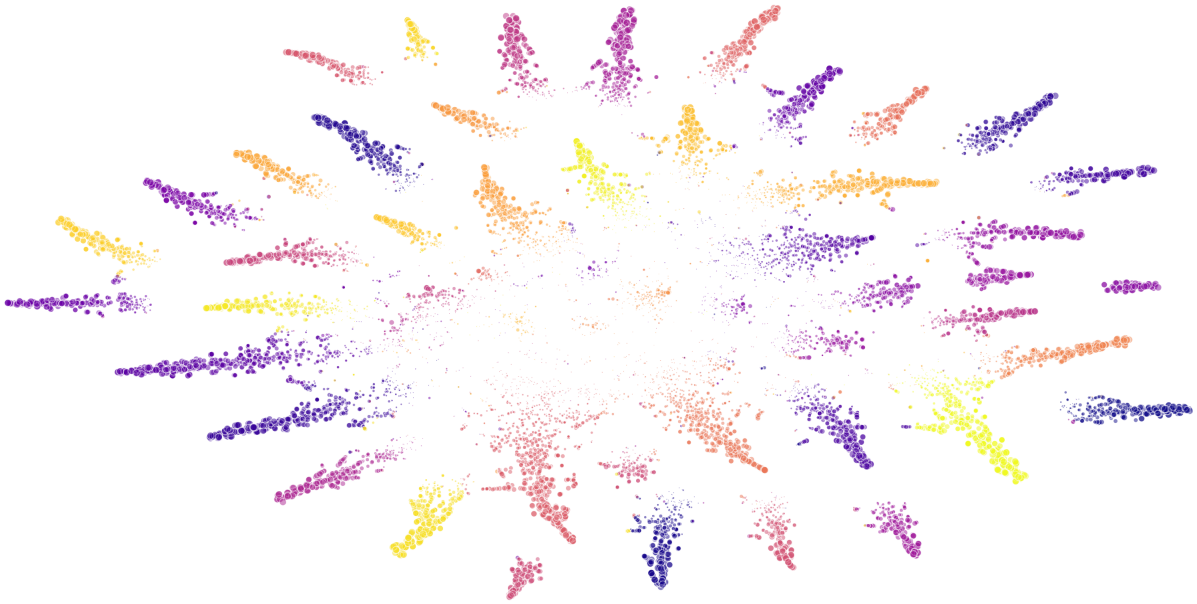
Topic Label	**Top Words**
Welfare and Unemployment	food stamp, unemployment benefit, welfare reform, money borrow, poor people, unemployment
Military Weapons	missile, ballistic missile, ballistic, submarine, silo, cruiser
Education	pupil, school district, private school, vocational education, school, local school district
Transportation and Infrastructure	passenger, airport, highway trust fund, highway trust, rail, mail service
Human Rights	captive nation, polish people, genocide, human right, human right violation, anniversary independence
Health Insurance	private insurance, health insurance, health plan, insurance company, premium, hospital care
Budget and Appropriations	budget authority outlay, authority outlay, budget authority, amount recommend, budget estimate
Land Entitlement	vessel, act entitle act, certain land, act entitle, subdivision, entitle act
Condolences and Tribute	condolence, mourn, untimely death, eulogy, passing, sense humor
Water and Navigation	navigable, nuclear waste, dam, flood control, basin, navigation
Judiciary and Trials	jury trial, jury, trial jury, grand jury, district judge, defendant
Trade and Industry	build ship, freedom independence, coastwise, farmer buy, cut taxis, coal miner
Scientific Research	space station, space, shuttle, scientist engineer, basic research, library
Postal Services	cent pound cent, pound cent, cent pound, postal employee, pound, advertising
Pensions and Social Security	grant pension, pension, increase pension, widow, social security benefit, retirement system
Banking and Finance	commercial bank, bank, national bank, silver dollar, loan association, depositor
Banking Regulation	hold company, national bank, banking, bank bank, bank, loan association
Foreign Aid	foreign aid, military assistance, military aid, military budget, economic aid, aid
Agriculture and Farming	wheat, bushel, feed grain, cent bushel, percent parity, bale
Elections and Voting	elector, ballot, vacancy, suffrage, legislative power, fill vacancy
Tariffs and Public Service	protective tariff, law degree, career public, pastor, treasurer, farm organization
Taxation	sale tax, property tax, tax income, property taxes, tax return, local taxes
Trade and Textile	trade agreement, textile, fair trade, free trade, textile industry, export
Public Buildings and Construction	circuit judge, cubic, public building, lieutenant, cadet, erection
Military Service and Veterans	active duty, veteran, disabled veteran, enlisted, reservist, enlisted man
Intelligence and Investigation	subpoena, grand jury, impeachment, investigation, intelligence agency, intelligence
Campaign Finance	spending limit, campaign finance, electoral, finance reform, candidate, campaign finance reform

Table 4: List of Estimated Topics from the *U.S. Congressional Record* (continued)

Topic Label	**Top Words**
Labor Market	wage rate, minimum wage, increase minimum wage, wage worker, cent hour, farm labor
Nuclear Treaties	nuclear test, nuclear arm, treaty, arm control, nuclear weapon, treaty treaty
Oil and Energy	domestic oil, barrel day, barrel oil, crude oil, oil, barrel
Public Health Risks	quarantine, ordnance, beverage, follow resolve, act entitle act, peace security
National Forests	national forest, wilderness, acre, wilderness area, grazing, timber
Legislative Process	unfinished business, divide control, general appropriation, unfinished, open hour, complete action
Energy Policy	change exist law, free election, national energy policy, testimony hearing, change exist, debatable
Capital Gains and Taxes	capital gain, gain tax, capital gain tax, tax rate, surtax, estate tax
Military Construction	military assistance, military construction, new project, increase pension, complete project, river harbor
Judicial and Highway Funding	district attorney, highway trust fund, highway trust, judgeship, parking, controller
Fisheries and Maritime Activities	fishery, coastwise, merchant marine, vessel, fishing, canal
Legal Claims	claimant, pay claim, claim claim, attorney fee, statute limitation, claim pay
Public Health	home health, disease, heart disease, mental health, health center, dental
Housing and Urban Development	affordable housing, public housing, urban renewal, housing unit, housing, voucher
Law Enforcement	law enforcement officer, enforcement officer, absent, pair, prisoner war, enlisted man
Forest Conservation	forest, rainfall, pest, timber, insect, acre
Natural Disasters	earthquake, evacuate, parking, evacuation, flooding, fire department
Crime and Justice	juvenile, crime, deport, firearm, violent crime, handgun
Intellectual Property and Commerce	copyright, patent, common carrier, interstate commerce, plaintiff, antitrust
War Negotiations	military aid, aggression, embargo, peace process, troop, peace
Sports and Social Issues	championship, colored, basketball, colored people, abortion, coach
Agricultural Commodities	butter, wool, pesticide, cent pound, cent pound cent, cent ad
Debt and Budget	debt limit, increase debt, public debt, balance budget, debt increase, interest debt

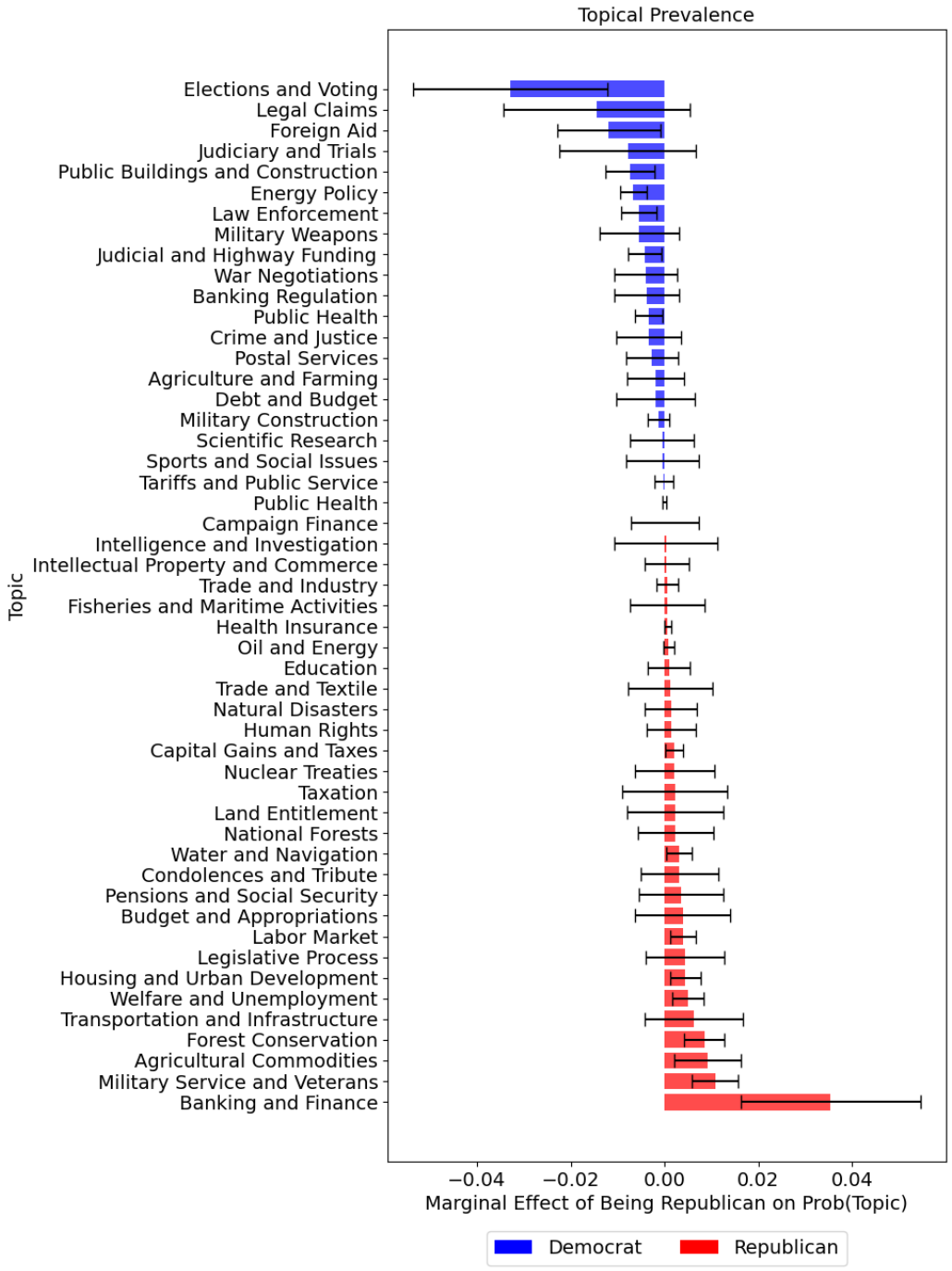
Notes: The 50 estimated GTM topics from the *U.S. Congressional Record*. Labels are manually assigned based on the top words. The top words are the most likely words to appear in a document that is 100% about said topic.

Figure 8: Two-Dimensional Projection of Speeches via t-SNE



Notes: 2-Dimensional projection via t-SNE of 20,000 randomly selected speeches from the *U.S. Congressional Record* represented as vectors of topic shares. The GTM creates soft clusters of the documents which are then interpreted as topics. Each dot represents a document. Colors represent the main topic for each document. Larger dots represent documents with a larger topic share of said topic.

Figure 9: Partisan Topics in the *U.S. Congressional Record*



Notes: Estimates of the coefficient in $\boldsymbol{\Gamma}$ associated with being Republican from the Logistic Normal prior. 95% confidence intervals computed with the parametric bootstrap. In blue, typically Democrat topics. In red, typically Republican topics.

Table 5: Content Associated with Covariates: Distinctive Words Over Time

Year of Congress	Top Words
Intercept	country, people, great, man, law, give
1917	conscription, prosecution war, win war, war measure, profiteer, present war
1919	profiteer, league, covenant, armistice, federal control, packer
1921	profiteer, armistice, prewar, bloc, tile, criticize
1929	stabilization, bus, care situation, human suffering, nitrogen, psychology
1931	moratorium, sale tax, commodity price, balancing, furlough, farm organization
1939	conscription, peacetime, national income, totalitarian, setup, airplane
1941	war effort, nondefense, defense industry, sabotage, defense effort, rubber
1943	war effort, postwar, parity price, win war, directive, print record part
1945	decontrol, atomic bomb, postwar, war effort, atomic energy, full employment
1947	atomic energy, atomic bomb, atomic, free enterprise, decontrol, ideology
1973	energy crisis, mass transit, legal service, energy problem, impact statement, environmental
1975	energy problem, service job, energy conservation, impact statement, food stamp, multinational
1977	energy problem, oversight hearing, economic stimulus, cruise missile, rate inflation, move opposition
1979	energy problem, oversight hearing, windfall profit, food stamp, inflation rate, auto industry
1981	regulatory reform, block grant, social security system, oversight hearing, auto industry, deferral
1993	health care reform, care reform, unfunded mandate, space station, deficit reduction, stimulus package
2001	war terrorism, homeland security, terrorist attack, economic stimulus, stimulus package, weapon mass destruction
2003	homeland security, war terror, war terrorism, terrorist attack, weapon mass destruction, terrorist
2005	war terror, homeland security, web site, border security, war terrorism, transparency
2007	transparency, war terror, homeland security, ensure, work bipartisan, urge reserve balance
2009	transparency, ensure, web site, stakeholder, job creation, health care reform
2011	transparency, job creation, job growth, infrastructure, ensure, bipartisan agreement

Notes: "Top words" gives the list of most likely words to appear in a document with this covariate—that is, documents from this biennium.

Table 6: Topic-Covariate Interactions: Party-Specific Topic Language

Topic Label	Type	Top Words
Welfare and Unemployment	Baseline	food stamp, unemployment benefit, welfare reform, money borrow, poor people, unemployment
	Republicans	money, man, pay, people, pension, come
	Democrats	unemployment, poor people, poverty, hungry, food stamp, unemployed
Elections and Voting	Baseline	elector, ballot, vacancy, suffrage, legislative power, fill vacancy
	Republicans	precinct, election, contestant, elector, ballot, seat
	Democrats	civil right, racial, equal protection, segregation, equal protection law, filibuster
Military Service and Veterans	Baseline	active duty, veteran, disabled veteran, enlisted, reservist, enlisted man
	Republicans	officer, man, soldier, enlistment, regiment, enlisted man
	Democrats	veteran, disabled veteran, personnel, care veteran, veteran family, veteran benefit
Campaign Finance	Baseline	spending limit, campaign finance, electoral, finance reform, candidate, campaign finance reform
	Republicans	election, candidate, party, money, postage, expenditure
	Democrats	campaign finance, spending limit, finance reform, campaign finance reform, campaign contribution, election campaign
Labor Market	Baseline	wage rate, minimum wage, increase minimum wage, wage worker, cent hour, farm labor
	Republicans	laborer, labor, man, railroad, workman, railway
	Democrats	minimum wage, wage worker, worker, increase minimum wage, cent hour, wage
Capital Gains and Taxes	Baseline	capital gain, gain tax, capital gain tax, tax rate, surtax, estate tax
	Republicans	taxation, cent, capital, taxis, pay, business
	Democrats	capital gain, income, tax, tax cut, income tax, excise
Legal Claims	Baseline	claimant, pay claim, claim claim, attorney fee, statute limitation, claim pay
	Republicans	claimant, claim, patent, treaty, property, land
	Democrats	insurance, insurance company, borrower, damage, mortgage, loan
Public Health	Baseline	home health, disease, heart disease, mental health, health center, dental
	Republicans	cord, care, service, hospital, physician, blood
	Democrats	health, mental health, health service, heart disease, cancer, public health
Crime and Justice	Baseline	juvenile, crime, deport, firearm, violent crime, handgun
	Republicans	officer, liquor, alien, consular, gun, intoxicate
	Democrats	juvenile, law enforcement, drug, enforcement, crime, criminal justice

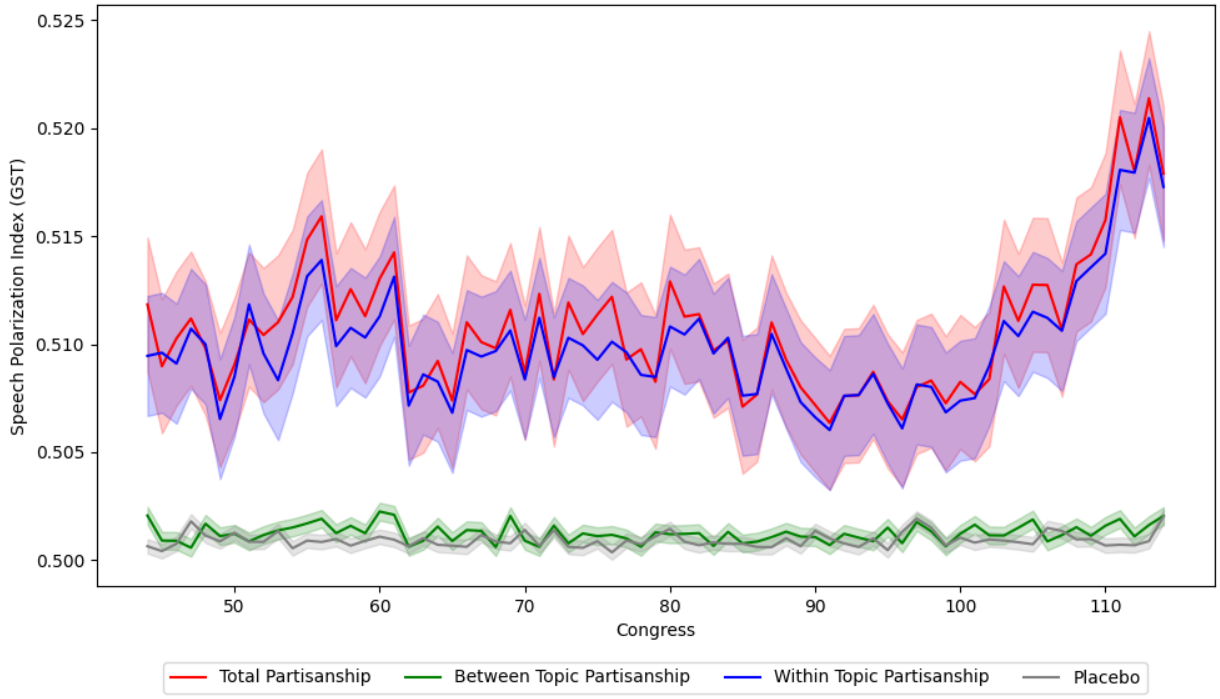
Notes: The 50 estimated topics. "Baseline" corresponds to most likely words for a document 100% about said topic. "Democrat" ("Republican") presents words whose probabilities change the most when interacting the topic with the content covariate for Democrats (Republicans).

Table 7: Quality of the Estimated Word Embeddings

Most Similar Phrases in Embeddings Space for:			
taxation	**immigration**	**war**	**health**
levy tax	immigration reform	military	health care
burden taxation	immigration law	troop	medical
taxes impose	immigrant come	soldier	care
tax income	illegal immigration	war war	disease
tax power	deport	country	cancer
tax cent	immigrant	peace	drug
amount revenue	undocumented	man	hospital
taxes levy	deportation	nation	child
wage	**election**	**regulation**	**fiscal**
minimum wage	election hold	law	budget
wage increase	ballot	regulate	appropriation
increase minimum wage	election election	regulatory	fund
increase minimum	election law	federal regulation	fiscal fiscal
raise minimum	day election	propose regulation	authorization
wage worker	candidate	new regulation	increase
labor	general election	law regulation	expenditure
wage hour	voter	company	fiscal budget
debt	**deficit**	**campaign**	**insurance**
national debt	deficit reduction	candidate	insurance company
public debt	budget deficit	campaign contribution	private insurance
debt debt	reduce deficit	campaign finance	insurance industry
pay debt	deficit problem	election	insurance coverage
debt pay	cut deficit	presidential campaign	insurance premium
debt limit	deficit deficit	election campaign	insurer
increase debt	federal deficit	political campaign	premium
pay interest	increase deficit	editorial	insured
recession	**growth**	**welfare**	**crime**
recession	economic growth	welfare reform	fight crime
unemployment benefit	growth rate	welfare system	violent crime
unemployed worker	growth economy	people welfare	commit crime
economic stimulus	rate growth	welfare people	crime commit
people unemployed	economic	welfare country	gun control
jobless	job creation	dependency	criminal justice
unemployment percent	economy grow	child care	crime rate
unemployment insurance	capital	public welfare	murder

Notes: Most similar phrases to specific words in embeddings space. The embeddings space we consider corresponds to the hidden layer of 256 units in the decoder. We measure similarity by cosine similarity. Overall, the GTM learns intuitive and interpretable word embeddings, where words with similar meanings are closer to one another in the embeddings space.

Figure 10: Trends in Speech Polarization in the US Congress (1873—2016)



Notes: Estimates of speech polarization between Republicans and Democrats for each session of Congress. The red line is estimated by simulating speeches where partisanship affects topic choice and topic content. The blue line is estimated by simulating speeches where partisanship affects topic content but not topic choice. The green line is estimated by simulating speeches where partisanship affects topic choice but not topic content. The grey line is estimated by simulating speeches where partisanship does not affect topic choice and topic content. Standard errors are obtained via the parametric bootstrap.

5 Application 2: Predicting business cycles with business news

Several finance and macroeconomics papers have recently integrated topic shares in macroeconometric models (Larsen and Thorsrud, 2019; Thorsrud, 2020; Bybee et al., 2024). They proceed in two steps. They first estimate topic shares via LDA and then use these topic shares as inputs for another econometric model. Conceptually, the main weakness of two-step approaches is that the downstream prediction model is not encouraged to learn topics that are predictive of the target outcomes. In contrast, GTM is encouraged to learn topics that lead both to a low reconstruction loss and a low mean-squared error in prediction. We show that the GTM allows researchers to effectively trade off better predictive power with topic interpretability.

5.1 Data and methods

Data. We apply the GTM to the full-text of *Wall Street Journal* articles published from 1984 to 2012. The dataset is provided as article-level counts of unigrams and bigrams. We lowercase, drop all digits, stopwords, tokens with strictly less than three characters, and tokens that appear

in at least 1% or in less than 0.1% of documents. This leaves us with a vocabulary of 13,893 phrases. We use Industrial Production Growth as a macroeconomic indicator which we download from the Federal Reserve Economic Data (FRED). The variable is relatively volatile, so we take the log and use a three-month rolling average as our prediction target.

Estimation details. We specify 50 topics for all models. The encoder and decoder have one hidden layer of 256 dimensions. The predictor has no hidden layer and is a simple linear mapping from topics to the outcome. All hidden layers have rectified linear unit (ReLU) activations. For training, a random sample of 50,000 training documents is selected with 50,000 documents reserved for validation. To prevent overfitting, we include a dropout rate of 20% while training. We stop updating parameters when the validation loss fails to improve.

For comparison, we also implement two-stage approaches. We train a plain-vanilla Latent Dirichlet Allocation (LDA) model via Gibbs sampling. We then train a linear regression with an L1 penalty (LASSO) selected via cross-validation to predict the same macroeconomic indicators.

Discussion. Before presenting our main results, two remarks are in order. First, our supervised learning task is formulated at the document-level. That is, we attempt to predict the smoothed Log Industrial Production Growth in month t by observing one article published in month $t-1$, which we represent as a vector of topic shares. This is an extremely challenging prediction task, as most articles don't mention industrial growth, so we do not expect the model to have high predictive accuracy.

Second, we do not use any external information for the prediction task (such as previous values of the target outcome). We are aware that actual macroeconometric algorithms aggregate topic shares within periods and use additional covariates for nowcasting. We consider this simple setup because it allows us to illustrate the GTM by (i) limiting to the predictive power of texts, and (ii) highlighting the potential benefits of modeling topics and outcomes jointly for supervised learning tasks. Our tools and insights are easily portable to more sophisticated architectures such as autoregressive models.

5.2 Forecasting macroeconomic indicators

Table 8 presents the performance of different modeling strategies for predicting economic outcomes from news articles. The two-stage approach based on LDA and LASSO predicts smoothed Log Industrial Production Growth with an R-squared of 0.030 on the training set and 0.023 on the test set. The GTM improves predictive performance across the specifications. When putting heavy weight on the prediction loss objective (λ_1 in Equation **??**), test-set R^2 goes as high as 0.076, more than triple the explanatory power of LDA.

The predictive lift from GTM is further illustrated in Figure 11. In particular, in Panel C, we can see that lasso (red) fails to capture the negative swings in the target variable. GTM (in green) is different, as it much better captures these negative dips in production growth.

Table 8: Forecasting Performance and Topic Diversity Across Models

Approach	Prediction Loss Weight (λ_1)	R^2 Training Set	R^2 Test Set	Topic Diversity
LDA + LASSO	-	0.030	0.023	0.74
GTM	0	-	-	0.464
	10 000 000	0.659	0.025	0.372
	50 000 000	0.699	0.051	0.244
	100 000 000	0.626	0.066	0.248
	1 000 000 000	0.321	0.076	0.098

Notes: This table compares the forecasting performance and topic quality of various models. We measure forecasting performance with the R-squared on the training and testing sets. Topic diversity is a measure of topic quality, where higher scores suggest better topics.

Next, we assess topic quality. Topic diversity measures the overlap of phrases across topics in the top phrases associated with each topic. In general, higher values suggest better topics. As seen in the last column of Table 8, vanilla LDA is a proper topic model and displays a high value of topic diversity, suggesting the model captures a broad range of diverse topics. GTM gets worse topic quality than LDA. Further, as we increase the weight given on the prediction loss, we observe a progressive decline in GTM's topic diversity, suggesting higher predictive accuracy may come at the expense of topic interpretability.

As an alternative visual representation of topic quality, Figure 12 plots topic representations of documents on a two-dimensional space via t-SNE. In general, well-delimited and relatively small clusters suggest better topics. Overall, the visual results reflect the topic diversity numbers. For GTM, the document clusters appear less and less clearly as we increase the weight of the prediction loss.

Hence, when increasing the prediction loss weight further, the GTM sacrifices the learning of topics for higher predictive accuracy. Such an apparent trade-off suggests that a grid-search approach on λ_1 may be an effective strategy for researchers wishing to strike the right balance between predictive performance and model interpretability.

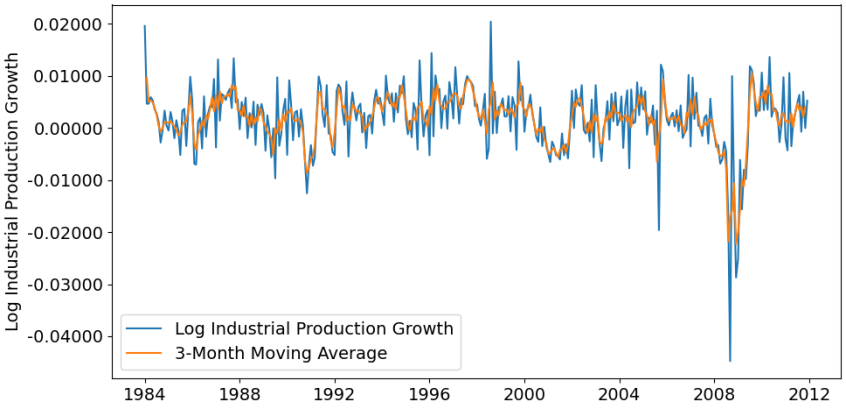
6 Application 3: An Image-Text topic model for political messaging on Facebook

This section shows the multimodal features of GTM by analyzing a joint corpus of text and images. Our data source is Meta's former public analysis tool CrowdTangle. We collected the universe of public Facebook posts by U.S. Congress from the years 2010 through 2022.

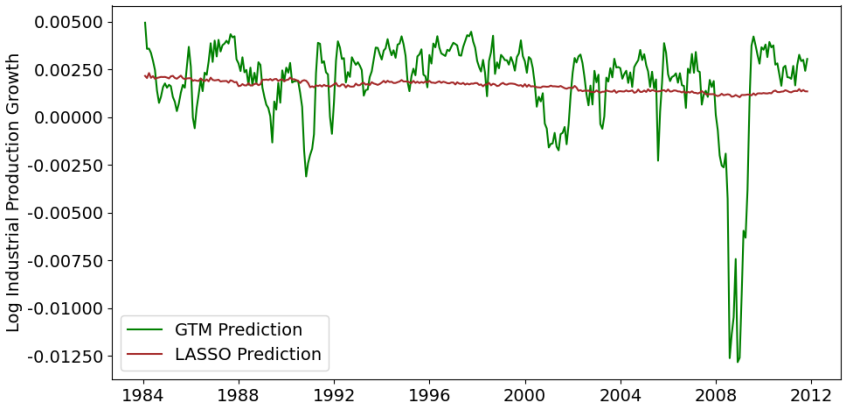
We analyze the set of posts with both an image and associated text snippet. We pre-process the data by removing posts with non-English text, removing hyperlinks from the texts, and keeping only verbs, nouns, and adjectives that are neither common English words nor part of a manually-created stop-word list to produce cleaned documents for the decoder. The final dataset consists of 553K image-text pairs.

Figure 11: Time Series of Forecasts for Log Industrial Production Growth

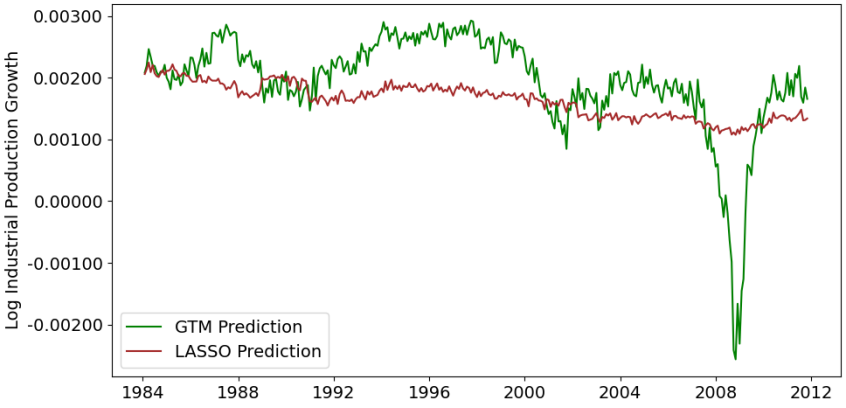
Panel A: Log Industrial Production Growth Over Time



Panel B: Predictions in the Training Set

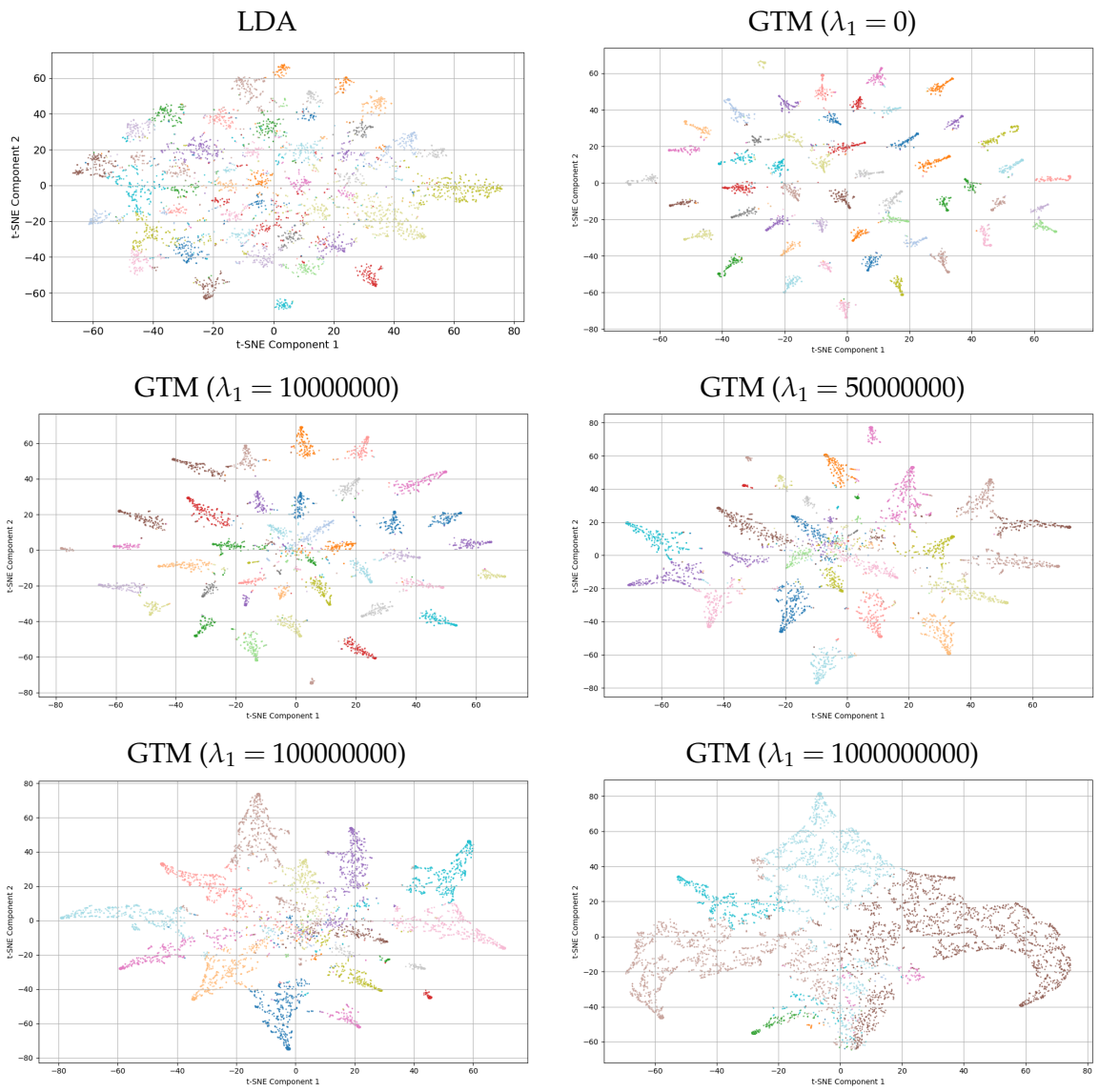


Panel C: Predictions in the Test Set



Notes: These figures compares the true values of Log Industrial Production Growth and the predicted values from LDA + LASSO and GTM with $\lambda_1 = 50000000$. Panel A shows the true target variable and the the 3-month rolling average. Panels B and C compare the predictions of LDA + LASSO (red) and GTM (green) for the training set (Panel B) and test set (Panel C).

Figure 12: Two-dimensional Visualization of the *Wall Street Journal* Articles via t-SNE



Notes: These figures present the resulting document clusters resulting from the estimation of LDA or the GTM with different weights assigned to the supervised prediction loss. The two dimensions are obtained via t-SNE. Each dot represents a document. Documents are colored based on their main topic. The document clusters appear less and less clearly as we increase the weight of the prediction loss λ_1.

We embed the texts and images using the multimodal encoder Long-CLIP (Zhang et al., 2024). CLIP is a deep learning model that learns a multimodal representation by training image-text pairs, optimizing the objective of predicting the matching image given the text and vice versa. *Long*-CLIP is an extension that allows for longer context lengths for the associated text documents, allowing up to 128 tokens. Luckily, the documents in our corpus are short and after cleaning, less than 1% of documents are more than 128 tokens.

We experimented with several GTM models using the CLIP embeddings, systematically varying the prior, the number of topics, the depth and width of the coder and decoder, and the content and prevalence covariates. The quality of the fit was assessed by the coherence of the top words per topic and the top documents. For the analysis below, we used a model with twenty topics, a hidden layer of 256 dimensions for both the encoder and decoder, a Dirichlet prior of one, and including covariates (all fixed effects) for time (year-month), state, and politician.

The list of top words associated with each of the 20 topics are shown in Table 9. The first column includes a topic label assigned by use based on the top-ranked topic words, top-ranked documents, and top-ranked images. Figure 13 shows representative images for each of the 20 topics (in the same order as the table). Overall, we see coherent topics that describe how Congress Members spend their time and how they communicate on social media. There are a number of topics about interacting with constituents, including celebrating holidays, get-out-the-vote, tributes to the military, and supporting local businesses. Further, there are a number of interesting policy topics, including guns, public health, environment, energy, women's rights, taxes, the economy, schools, and immigration. Both the words and images are intuitively related to the specified topic label.

How prevalent are these topics in the texts and images? Table 10 gives the share of topics that occur in text (third column) and images (fourth column). The shares are quite evenly distributed with topics taking up between 1% and 10% of the posts for both modalities. In the fifth column, we show the percent difference of the image share relative to the text share, and this is also the sorting metric. Interestingly, somewhat emotive and divisive topics such as gun violence, immigration, and women's rights, and schools are featured in images. Instead, relatively dry topics such as public health, taxes, infrastructure, small businesses, and the economy are communicated through text.

Now we analyze how topics are related to partisanship. Figure 14 gives a scatter plot of the 20 topics, with each mark labeled. The horizontal axis is the left-right spectrum going from topics that are used more by Democrats in the texts to topics that are used more by Republicans in the texts. Correspondingly, the vertical axis is the image slant, going from more Republican at the top to more Democrat at the bottom. We can see that, overall, the text and image slant of topics is highly correlated. Overall, the level of slant is similar in the text and images, with text documents slightly more slanted. We don't see any topics that are more slanted among images except the holidays and awards topic, which tends to be more slanted toward Republicans in the image space.

Finally, we look at whether the topics in Facebook posts became more polarized over the 2010-2022 time period. As a simple descriptive exercise, we ran logistic regressions to predict the politi-

Table 9: Joint Text-Image Topics in Congress Member Facebook Posts

Topic	Label	Representative Words
1	Sports and Competitions	game, win, luck, fun, team, young, football, congrat
2	Holidays and Awards	wish, family, present, love, dad, mom, hope, free
3	Voting System	vote, voting, election, democracy, state, abortion, fight, ballot
4	Military Honors	present, woman, man, man woman, brave, medal, hero, military
5	Gun Violence	gun, gun violence, violence, border, statement, vote, action, shooting
6	COVID-19 Health Services	vaccine, site, free, test, health, 19, covid, covid 19
7	Natural Disasters	storm, fire, update, damage, evacuation, stay, stay safe, area
8	Clean Energy & Infrastructure	energy, infrastructure, clean, create, project, water, bridge, economy
9	Campaign Volunteering & Outreach	volunteer, check, constituent, campaign, update, hope, lot, folk
10	Tragedy & Mourning	lose, forget, world, love, fight, man, prayer, attack
11	Women's Rights	woman, fight, equality, equal, vote, earn, legacy, woman vote
12	Voting Encouragement	vote, voting, early, polling, poll, early voting, ballot, location
13	Taxes	tax, plan, health, cost, taxis, insurance, tax cut, bill
14	Business Openings	facility, grand, manufacturing, economy, opening, company, grow, grand opening
15	Public Health	fire, damage, funding, health, flood, recovery, project, flooding
16	Military Sacrifice & Remembrance	man, brave, man woman, woman, ultimate, ultimate sacrifice, forget, hero
17	Economy	tax, american, economy, growth, create, code, energy, cost
18	Farming & Small Business	farmer, facility, manufacturing, economy, company, industry, product, grow
19	Schools	teacher, school, class, young, education, bright, app, classroom
20	Immigration & Border Control	border, illegal, southern border, southern, illegal immigrant, statement, immigrant, crisis

Figure 13: Top Images for each of the 20 Topics

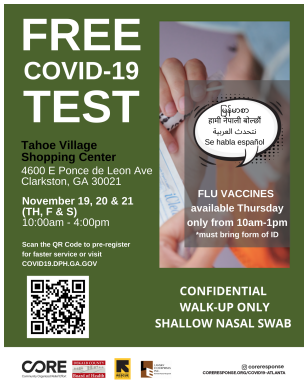








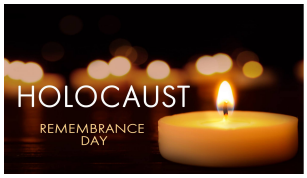


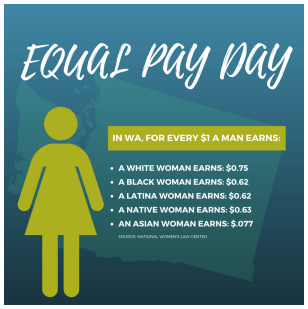














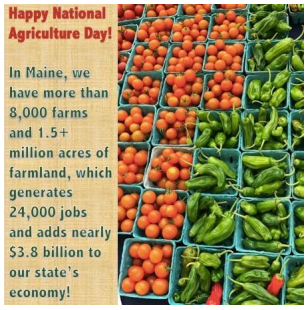














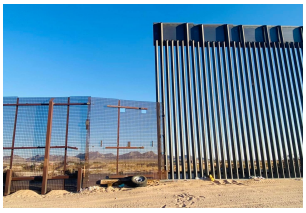


Table 10: Topic Prevalence in Text and Images

Topic Code	Topic Label	Prevalence (Text)	Prevalence (Image)	Image – Text (%)
5	Gun Violence	0.045	0.081	80.30
2	Holidays and Awards	0.047	0.084	77.27
3	Voting System	0.023	0.039	71.94
10	Tragedy & Mourning	0.041	0.066	61.10
20	Immigration & Border Control	0.039	0.060	53.90
11	Women's Rights	0.051	0.074	45.10
12	Voting Encouragement	0.061	0.083	36.68
19	Schools	0.084	0.106	25.99
4	Military Honors	0.055	0.050	-9.45
9	Campaign Volunteering & Outreach	0.080	0.070	-12.20
6	COVID-19 Health Services	0.053	0.046	-12.76
7	Natural Disasters	0.051	0.044	-13.99
1	Sports and Competitions	0.060	0.051	-14.66
16	Military Sacrifice & Remembrance	0.041	0.034	-15.53
14	Business Openings	0.042	0.030	-30.04
17	Economy	0.050	0.022	-55.54
18	Farming & Small Business	0.046	0.017	-61.83
8	Clean Energy & Infrastructure	0.048	0.018	-62.06
13	Taxes	0.045	0.015	-65.42
15	Public Health	0.039	0.008	-79.01

Figure 14: Slant by Topic for Texts and Images

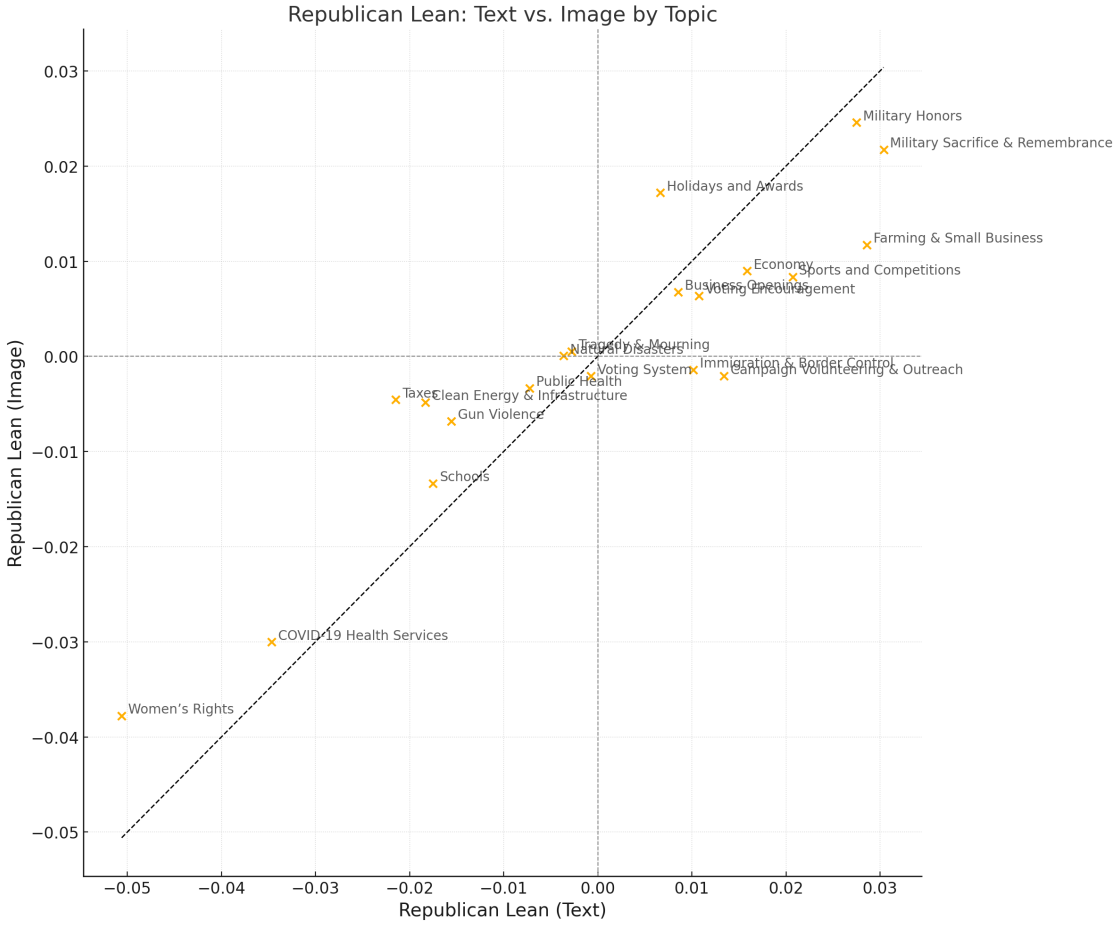
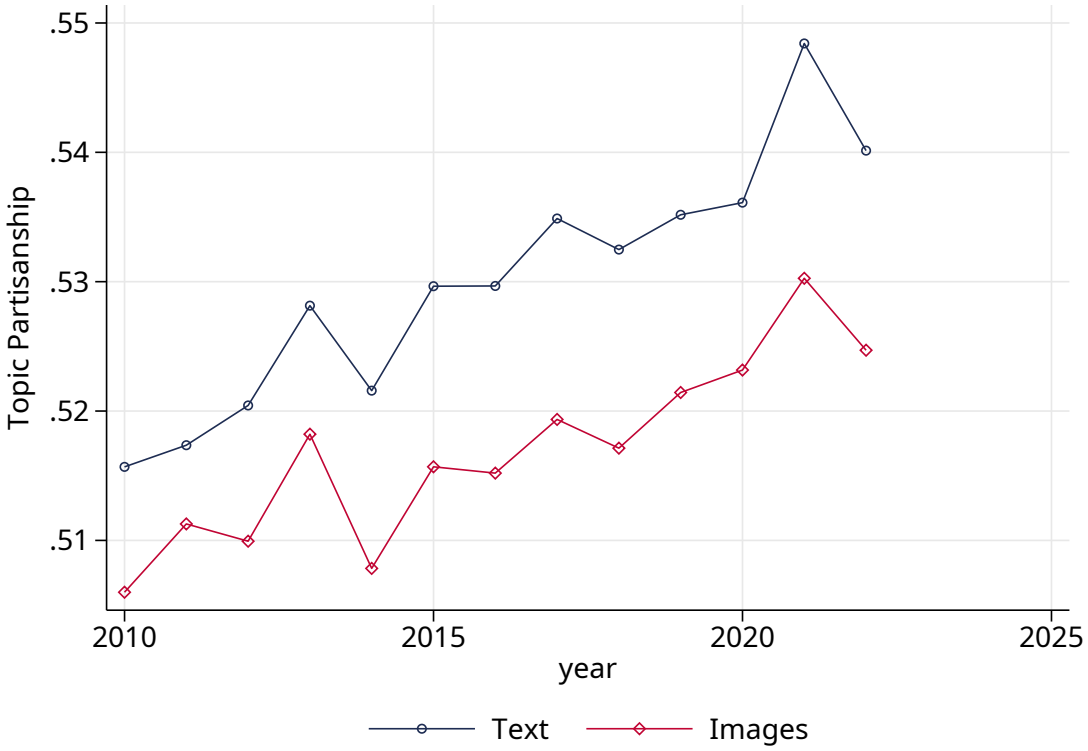


Figure 15: Topic-Predicted Slant Over Time for Texts and Images



cal party of a congressman from the topic shares, separately for texts and images. We then formed the predicted values and plotted the share of "true" predictions over time, which gives an idea of how well someone can guess party from the topics covered in texts and images.

These predicted values over time are shown in figure is shown in Figure 15. As can be seen, topics are predictive of party and that predictiveness increases over the time period, for both text (in blue) and images (in red). The predictiveness of text topics for party has increased relatively more. That suggests that increasing political polarization over this time period is through language rather than imagery.

7 Conclusion

This paper develops a general econometric framework for deep latent factor models that can accommodate high-dimensional, unstructured, and multimodal measurements. We specify a modular data-generating process in which a low-dimensional latent factor generates multiple observed modalities and auxiliary outcomes, conditional on covariates that can shift prevalence, content, and supervision channels. To estimate these models at scale, we use amortized variational inference and maximize an ELBO on the marginal likelihood, with a training strategy designed to remain well-defined when different units have different observed modality sets.

We clarify the statistical content of ELBO-based estimation by characterizing conditions un-

der which the inference gap can be made arbitrarily small, and by decomposing approximation error into variational and amortization components. In an extended linear–Gaussian benchmark, the posterior is available in closed form, the corrected product-of-experts fusion rule is exact, and ELBO maximization inherits familiar likelihood-based consistency and asymptotic normality results. Together, these results explain when amortized variational procedures can be interpreted as likelihood-based estimators and provide guidance on how modeling choices for priors, decoders, and variational families affect approximation error in practice.

Empirically, we illustrate the framework through three applications. In U.S. Congressional speech, the model supports a decomposition of speech polarization into topic choice and within-topic framing. In business news, jointly learning topics and prediction targets improves out-of-sample forecasting performance relative to a two-step LDA+LASSO pipeline, while revealing a tradeoff between predictive accuracy and topic interpretability. In multimodal political communication on Facebook, the model learns coherent joint text–image topics and documents differences in how topics are expressed across modalities and across parties. More broadly, the framework is designed to be portable to other social-science settings where the measurement problem is high-dimensional, multimodal, and only partially observed, while retaining a clear likelihood-based interpretation and transparent conditions for exact inference.

References

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., Fersini, E., et al. (2021). Cross-lingual contextualized topic models with zero-shot learning. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1676–1683. Association for Computational Linguistics (ACL).

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

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Józefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21. Association for Computational Linguistics.

Bybee, L., Kelly, B. T., Manela, A., and Xiu, D. (2024). Business news and business cycles. *Journal of Finance*.

Chakraborty, S. and Bartlett, P. (2024). A statistical analysis of wasserstein autoencoders for intrinsically low-dimensional data. In *The Twelfth International Conference on Learning Representations*.

Cremer, C., Li, X., and Duvenaud, D. (2018). Inference suboptimality in variational autoencoders. In *International conference on machine learning*, pages 1078–1086. PMLR.

Gatopoulos, I., Stol, M., and Tomczak, J. M. (2020). Super-resolution variational auto-encoders. *arXiv preprint arXiv:2006.05218*.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized contrastive learning. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29*, pages 4743–4751.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Larsen, V. H. and Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, 210(1):203–218.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. (2019). Don't blame the ELBO! a linear VAE perspective on posterior collapse. In *Advances in Neural Information Processing Systems (NeurIPS)*. Accepted at NeurIPS 2019.

Margossian, C. C. and Blei, D. M. (2024). Amortized variational inference: When and why? *Journal of the American Statistical Association*, 119(545):65–78.

Mcauliffe, J. and Blei, D. (2007). Supervised topic models. *Advances in neural information processing systems*, 20.

Moran, G. E., Sridhar, D., Wang, Y., and Blei, D. M. (2021). Identifiable deep generative models via sparse decoding. *arXiv preprint arXiv:2110.10804*.

Nan, F., Ding, R., Nallapati, R., and Xiang, B. (2019). Topic modeling with wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *ICML'15*, pages 1530–1538. JMLR.org.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

Rodriguez, P. L. and Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1):101–115.

Shi, Y., Paige, B., Torr, P., et al. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Srivastava, A. and Sutton, C. (2022). Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Thorsrud, L. A. (2020). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 38(2):393–409.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2018). Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Tomczak, J. M. and Welling, M. (2018). Vae with a vampprior. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1214–1223, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Wang, Y., Ainsworth, S. K., and Blei, D. M. (2023). Posterior collapse and identifiability in variational autoencoders. *arXiv preprint arXiv:2301.00537*.

Wu, M. and Goodman, N. (2018). Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31.

Zhang, B., Zhang, P., Dong, X., Zang, Y., and Wang, J. (2024). Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

A Notation

Table 11: Summary of notation

Symbol	Description
Indices and dimensions	
$i \in \{1, \dots, n\}$	Unit (observation) index
$m \in \{1, \dots, M\}$	Modality index
$k \in \{1, \dots, K\}$	Auxiliary outcome index
L	Dimension of the latent space
V	Vocabulary size (for text data)
D	Dimension of observed data (e.g., in PPCA)
Observed variables	
w_{im}	Response variable for unit i, modality m
$w_{i,S}$	Response variables for unit i, modality subset S
y_{ik}	Auxiliary outcome k for unit i
x_i	Covariates for unit i
x_i^p	Prevalence covariates (shift prior distribution)
x_i^c	Content covariates (shift modality likelihoods)
x_i^s	Supervision covariates (shift outcome likelihoods)
$d_{i,S}$	Observed data: $(w_{i,S}, y_{i,1:K})$
O_i	Set of observed modalities for unit i
S_i	Subset of modalities used for training/inference
Latent variables	
$z_i \in \mathcal{Z} \subseteq \mathbb{R}^L$	Latent factor vector for unit i
Model parameters	
ψ	Prior parameters
θ_m	Decoder parameters for modality m
θ_k	Predictor parameters for outcome k
$\theta = \{\theta_m, \theta_k\}$	All decoder/predictor parameters
ϕ	Encoder (variational) parameters
$\vartheta = (\theta, \psi)$	All generative model parameters
Distributions	
$p_\psi(z_i \mid x_i^p)$	Prior distribution over latent factors
$p_{\theta_m}(w_{im} \mid z_i, x_i^c)$	Likelihood for modality m
$p_{\theta_k}(y_{ik} \mid z_i, x_i^s)$	Likelihood for auxiliary outcome k
$q_\phi(z_i \mid \cdot)$	Variational (approximate posterior) distribution
$q_\phi^{(m)}(z_i \mid w_{im}, \cdot)$	Modality-specific encoder
$r(\cdot \mid O_i)$	Modality subsampling distribution
Inference quantities	
$\mathcal{L}_i(S; q)$	Evidence lower bound (ELBO) for unit i, subset S
$\mathcal{L}_i^\star(S)$	Optimal ELBO within variational family $\mathcal{Q}$
$\mathcal{J}(\theta, \phi, \psi)$	Training objective (sum of expected ELBOs)
$\mathrm{KL}(q \Vert p)$	Kullback–Leibler divergence from p to q
Weighted ELBO hyperparameters	
β	Weight on KL divergence term
γ_k	Weight on auxiliary outcome k

B Special cases and notation mapping

This appendix records several common latent factor models as special cases of the data-generating process in Eq. (1). Throughout, we keep the notation $(z_i, x_i^p, x_i^c, x_i^s, w_{im}, y_{ik})$ fixed and specify how each special case sets M, K, the prior $p_\psi(\cdot)$, and the conditional distributions $p_{\theta_m}(\cdot)$ and $p_{\theta_k}(\cdot)$.

B.1 Probabilistic PCA (PPCA)

Probabilistic PCA (Tipping and Bishop, 1999) is a classical latent factor model for continuous data. It assumes that high-dimensional observations are generated by a linear transformation of low-dimensional latent factors plus Gaussian noise. Set $M = 1$, $K = 0$, x_i^p degenerate (or empty), and denote $w_{i1} \equiv w_i \in \mathbb{R}^D$. Let $z_i \in \mathbb{R}^L$ with $L < D$.

$$
\begin{aligned}
z_i &\sim p_\psi(z_i \mid x_i^p) = \mathcal{N}(0, I_L), \\
w_{i1} &\sim p_{\theta_1}(w_{i1} \mid z_i, x_i^c) = \mathcal{N}(Wz_i + \mu, \sigma^2 I_D),
\end{aligned}
$$

where $\theta_1 = (W, \mu, \sigma^2)$ consists of a $D \times L$ loading matrix W, a mean vector $\mu \in \mathbb{R}^D$, and a noise variance $\sigma^2 > 0$.

B.2 Structural Topic Model

The Structural Topic Model (Roberts et al., 2016) extends latent Dirichlet allocation by allowing document-level covariates to influence topic prevalence and topic content. It is widely used in political science and computational social science. Set $M = 1$, $K = 0$, and let $w_{i1} \equiv w_i$ be the document word-count vector of length V (vocabulary size) with total word count $N_i = \sum_{v=1}^{V} w_{iv}$. Let the prevalence covariates be $x_i^p \equiv X_i$. The STM uses a logistic-normal prior on topic proportions:

$$
\begin{aligned}
z_i &\sim p_\psi(z_i \mid x_i^p) = \mathrm{LN}(\Gamma' X_i, \Sigma), \\
w_{i1} &\sim p_{\theta_1}(w_{i1} \mid z_i, x_i^c) = \mathrm{Multinomial}\big(N_i,\ f_{\theta_1}(z_i, x_i^c)\big),
\end{aligned}
$$

where $\mathrm{LN}(\cdot, \cdot)$ denotes the logistic-normal distribution, Γ is a matrix of prevalence coefficients, Σ is the topic covariance, and $f_{\theta_1}(z_i, x_i^c)$ returns word probabilities in the simplex Δ^{V-1}.

B.3 Supervised Topic Model

The Supervised Topic Model (Mcauliffe and Blei, 2007) augments a topic model with a prediction task: the latent topic proportions are used to predict an auxiliary outcome such as a document rating or label. Set $M = 1$, $K = 1$, let $w_{i1} \equiv w_i$ denote the document word counts, and let $y_{i1} \equiv y_i$

be an auxiliary outcome.

$$
\begin{aligned}
z_i &\sim p_\psi(z_i \mid x_i^p) = \mathrm{LN}(0, I), \\
w_{i1} &\sim p_{\theta_1}(w_{i1} \mid z_i, x_i^c) = \mathrm{Multinomial}(N_i, \Phi z_i), \\
y_{i1} &\sim p_{\theta_1^{(y)}}(y_{i1} \mid z_i, x_i^s),
\end{aligned}
$$

where Φ is a $V \times L$ matrix of topic–word probabilities and the outcome distribution $p_{\theta_1^{(y)}}$ is task-dependent.

B.4 Multimodal Image–Text Model

Multimodal VAEs learn a shared latent representation from multiple data modalities, such as images and their captions. The key idea is that the same latent factors should explain variation in both modalities. Set $M = 2$, with w_{i1} the text modality (e.g., caption word counts or embeddings) and w_{i2} the image modality (e.g., pixel values or CNN features). The model is:

$$
\begin{aligned}
z_i &\sim p_\psi(z_i \mid x_i^p), \\
w_{i1} &\sim p_{\theta_1}(w_{i1} \mid z_i, x_i^c), \\
w_{i2} &\sim p_{\theta_2}(w_{i2} \mid z_i, x_i^c),
\end{aligned}
$$

where p_{θ_1} and p_{θ_2} are instantiated using likelihoods and decoder architectures appropriate to each modality. For text, p_{θ_1} might be a categorical distribution over words decoded by an LSTM or Transformer; for images, p_{θ_2} might be a Gaussian distribution over pixels decoded by a convolutional neural network.

C An alternative model: The Wasserstein Autoencoder (WAE)

The Wasserstein autoencoder (WAE) addresses variational inference from an optimal transport perspective (Tolstikhin et al., 2018). The Wasserstein distance between the data distribution P_W and the generative model distribution P_G can be written as

$$
W_c(P_W, P_G) = \inf_{\Gamma \in \Pi(P_W, P_G)} \mathbb{E}_{(W, \hat{W}) \sim \Gamma}[c(W, \hat{W})],
$$

where $\Pi(P_W, P_G)$ is the set of couplings with marginals P_W and P_G. Direct minimization of $W_c(P_W, P_G)$ is intractable in high dimensions. Introducing an encoder distribution $Q(Z|W)$ and a decoder $G(Z)$ yields an upper bound to the transport cost:

$$
W_c(P_W, P_G) \leq \mathbb{E}_{P_W} \mathbb{E}_{Q(Z|W)}[c(W, G(Z))] + D(Q_Z, P_Z).
$$

This reformulation reveals the WAE objective as a *penalized optimal transport* (POT) problem. The first term corresponds to the cost of transporting observed data W through the encoder–decoder

pipeline, while the second term enforces that the aggregate posterior Q_Z matches the prior P_Z. This differs from the variational autoencoder (VAE), where $Q(Z|W)$ is regularized against P_Z for each data point, often leading to over-constrained encodings. In contrast, the WAE only enforces *aggregate* matching, which allows greater flexibility in the learned representations while still nudging them toward the prior.

Ultimately, we can recast the problem as a Lagrangian:

$$\inf_{G\in\mathcal{G},Q(Z|W)\in\mathcal{Q}} \mathbb{E}_{P_W}\mathbb{E}_{Q(Z|W)}[c(W,G(Z))] + \lambda D(Q_Z,P_Z) \tag{22}$$

where $c(\cdot)$ is a cost function,[2] $Q_Z = \mathbb{E}_{P_W} Q(Z|W)$ is the aggregated posterior, $D(Q_Z,P_Z)$ is a divergence metric between Q_Z and P_Z, and $0 \leq \lambda$ is a hyperparameter controlling the degree of regularization of the encodings distribution. Reminiscent of standard Bayesian inference, Equation (22) consists of a reconstruction term and a regularization term. The reconstruction term pushes the autoencoder to closely reconstruct the original target distribution P_W. The regularization term pushes the encoding distribution Q_Z to remain close to the prior distribution P_Z.

As a divergence metric $D(\cdot)$, Tolstikhin et al. (2018) propose the Maximum Mean Discrepancy (MMD). Let $\mathcal{H}$ denote the Reproducing Kernel Hilbert Space of real-valued functions mapping $\mathcal{Z}$ to $\mathbb{R}$ (Berlinet and Thomas-Agnan, 2011). For a kernel function $k : \mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$, the MMD is defined as:

$$\text{MMD}_k(Q_Z,P_Z) = || \int_{\mathcal{Z}} k(z,\cdot)dP_Z(z) - \int_{\mathcal{Z}} k(z,\cdot)dQ_Z(z) \,||_{\mathcal{H}_k},$$

and can be unbiasedly estimated using m samples drawn from Q_Z and P_Z:

$$\widehat{\text{MMD}}^2_k(x,y) = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k(x_i,x_j) + \frac{1}{m(m-1)}\sum_{i\neq j}^{n} k(y_i,y_j) - \frac{2}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(x_i,y_j).$$

Chakraborty and Bartlett (2024) provide statistical encoding and decoding guarantees for WAEs. Under mild regularity and smoothness assumptions, they prove that the estimated encoded distribution converges to the true latent distribution:

$$\frac{1}{n}\sum_{k=0}^{n} Q(Z|w_k) \xrightarrow{d} P_Z.$$

Furthermore, the generator can almost surely map back the encoded points to the original data as the number of observations grows to infinity. However, since the MMD loss matches the aggregate posterior, there is no guarantee that the model converges in probability. Despite these weaker theoretical guarantees, the WAE enjoys very stable training relative to its maximum likelihood counterparts, can accommodate any prior distribution, and its MMD divergence is fast to compute relative to other divergence metrics. It is also straightforward to extend it to multiple modalities, auxiliary outcomes, and covariates—as our main text model.

[2] In practice, if W is continuous, the reconstruction loss is often the mean squared error between the observation's representation W and the predicted value from the autoencoder network. If W is categorical, the reconstruction loss for the autoencoder is often the negative cross-entropy loss.

D Proofs

D.1 Derivation of the ELBO

We derive the evidence lower bound stated in Equation (9). Fix a unit i and a subset of modalities $S \subseteq \{1, \dots, M\}$. Starting from the log marginal likelihood, we introduce the variational distribution $q(z_i)$ via importance sampling and apply Jensen's inequality:

$$\begin{aligned}
\log p_{\theta,\psi}(w_{i,S}, y_{i,1:K} \mid x_i) &= \log \int p_\psi(z_i \mid x_i^p) \left[\prod_{m \in S} p_{\theta_m}(w_{im} \mid z_i, x_i^c) \right] \left[\prod_{k=1}^{K} p_{\theta_k}(y_{ik} \mid z_i, x_i^s) \right] dz_i \\
&= \log \int \frac{q(z_i)}{q(z_i)} \, p_\psi(z_i \mid x_i^p) \prod_{m \in S} p_{\theta_m}(w_{im} \mid z_i, x_i^c) \prod_{k=1}^{K} p_{\theta_k}(y_{ik} \mid z_i, x_i^s) \, dz_i \\
&= \log \mathbb{E}_q \left[\frac{p_\psi(z_i \mid x_i^p) \prod_{m \in S} p_{\theta_m}(w_{im} \mid z_i, x_i^c) \prod_{k=1}^{K} p_{\theta_k}(y_{ik} \mid z_i, x_i^s)}{q(z_i)} \right]
\end{aligned}$$

By Jensen's inequality (since log is concave),

$$\mathbb{E}_q \left[\log \frac{p_\psi(z_i \mid x_i^p) \prod_{m \in S} p_{\theta_m}(w_{im} \mid z_i, x_i^c) \prod_{k=1}^{K} p_{\theta_k}(y_{ik} \mid z_i, x_i^s)}{q(z_i)} \right] \;\le\; \log p_{\theta,\psi}(w_{i,S}, y_{i,1:K} \mid x_i).$$

Expanding the logarithm and rearranging terms yields

$$\begin{aligned}
\mathcal{L}_i(S; q) &:= \mathbb{E}_q \Big[\log p_\psi(z_i \mid x_i^p) + \sum_{m \in S} \log p_{\theta_m}(w_{im} \mid z_i, x_i^c) + \sum_{k=1}^{K} \log p_{\theta_k}(y_{ik} \mid z_i, x_i^s) - \log q(z_i) \Big] \\
&= \sum_{m \in S} \mathbb{E}_q \big[\log p_{\theta_m}(w_{im} \mid z_i, x_i^c) \big] + \sum_{k=1}^{K} \mathbb{E}_q \big[\log p_{\theta_k}(y_{ik} \mid z_i, x_i^s) \big] \\
&\quad - \mathrm{KL}\Big(q(z_i) \,\big\|\, p_\psi(z_i \mid x_i^p) \Big).
\end{aligned}$$

Combining the previous display with Jensen's inequality implies

$$\mathcal{L}_i(S; q) \;\le\; \log p_{\theta,\psi}(w_{i,S}, y_{i,1:K} \mid x_i).$$

Moreover, the gap between the log marginal likelihood and the ELBO equals the Kullback–Leibler divergence between the variational distribution and the true posterior:

$$\log p_{\theta,\psi}(w_{i,S}, y_{i,1:K} \mid x_i) = \mathcal{L}_i(S; q) + \mathrm{KL}\Big(q(z_i) \,\big\|\, p_{\theta,\psi}(z_i \mid w_{i,S}, y_{i,1:K}, x_i) \Big).$$

Since the Kullback–Leibler divergence is nonnegative, this confirms the lower-bound property and shows that the bound is tight if and only if $q(z_i)$ coincides with the true posterior almost surely.

D.2 Our model is a simple hierarchical model.

We verify that our model is a *simple hierarchical model* (SHM), and explain why SHMs satisfy condition (iii) in Proposition 1 under the regularity assumptions in Margossian and Blei (2024). The key observation is that our model can be written in the form

$$p(\vartheta)\prod_{i=1}^{n} p(z_i \mid \vartheta, x_i)\, p(d_{i,S} \mid z_i, \vartheta, x_i),$$

where $\vartheta := (\theta, \psi)$ collects all global parameters. This factorization shows that, conditional on the global parameters ϑ and covariates x_i, the local variables $(z_i, d_{i,S})$ are independent across units i. Models with this structure are called *simple hierarchical models* (SHMs). By Margossian and Blei (2024), simple hierarchical models admit an ideal inference function: there exists a mapping $\mathcal{I}_S^\star$ from each unit's data $(d_{i,S}, x_i)$ to variational parameters $\lambda_i^\star(S) \in \Lambda$ such that $q_{\lambda_i^\star(S)}$ achieves the variational optimum $\mathcal{L}_i^\star(S)$. Moreover, under mild regularity conditions, one can select (exact or δ-approximate) optimizers as continuous functions on compact domains, which yields the continuous near-optimal selector required by condition (iii).

In our notation, since $\mathcal{L}_i(S;q) = \log p_\vartheta(d_{i,S} \mid x_i) - \mathrm{KL}(q \| p_\vartheta(\cdot \mid d_{i,S}, x_i))$, maximizing the ELBO over $\mathcal{Q}$ is equivalent to minimizing $\mathrm{KL}(q \| p_\vartheta(\cdot \mid d_{i,S}, x_i))$ over $\mathcal{Q}$. Thus, continuity of an (approximate) ELBO optimizer yields exactly the continuous near-optimal selector required by condition (iii).

E Extended linear–Gaussian case: posterior and fusion algebra

This appendix records the closed-form subset posterior and the Gaussian fusion algebra used in Section 2.3. We work throughout with the extended conditional linear–Gaussian case (multimodal, covariates, continuous outcomes) and a fixed subset $S \subseteq \{1, \ldots, M\}$.

Subset posterior via completing the square. Recall the model:

$$\begin{aligned} z_i \mid x_i^p &\sim \mathcal{N}(\mu_0(x_i^p), \Sigma_0), \\ w_{im} \mid (z_i, x_i^c) &\sim \mathcal{N}(b_m(x_i^c) + A_m(x_i^c) z_i, \Sigma_{w,m}), \quad m \in S, \\ y_i := (y_{i1}, \ldots, y_{iK})^\top, \qquad y_i \mid (z_i, x_i^s) &\sim \mathcal{N}(d(x_i^s) + C(x_i^s) z_i, \Sigma_y), \end{aligned}$$

with conditional independence across m and between $(w_{i,S})$ and y_i given (z_i, x_i). Consistency with the factorization in Eq. (1) corresponds to taking Σ_y diagonal. Define residuals $\widetilde{w}_{im} := w_{im} - b_m(x_i^c)$

and $\widetilde{y}_i := y_i - d(x_i^s)$. Up to an additive constant independent of z_i, the conditional log posterior is

$$\begin{aligned}\log p_\vartheta(z_i \mid w_{i,S}, y_i, x_i) = &-\tfrac{1}{2}(z_i - \mu_0)^\top \Sigma_0^{-1}(z_i - \mu_0) \\ &- \tfrac{1}{2}\sum_{m\in S}(\widetilde{w}_{im} - A_m z_i)^\top \Sigma_{w,m}^{-1}(\widetilde{w}_{im} - A_m z_i) \\ &- \tfrac{1}{2}(\widetilde{y}_i - C z_i)^\top \Sigma_y^{-1}(\widetilde{y}_i - C z_i),\end{aligned}$$

where we suppress (x_i^p, x_i^c, x_i^s) in μ_0, A_m, C for readability. Expanding and collecting terms in z_i yields a quadratic form

$$\log p_\vartheta(z_i \mid w_{i,S}, y_i, x_i) = -\tfrac{1}{2} z_i^\top V_{i,S}^{-1} z_i + z_i^\top V_{i,S}^{-1} m_{i,S} + \text{const},$$

with

$$\begin{aligned}V_{i,S}^{-1} &= \Sigma_0^{-1} + \sum_{m\in S} A_m^\top \Sigma_{w,m}^{-1} A_m + C^\top \Sigma_y^{-1} C, \\ V_{i,S}^{-1} m_{i,S} &= \Sigma_0^{-1}\mu_0 + \sum_{m\in S} A_m^\top \Sigma_{w,m}^{-1} \widetilde{w}_{im} + C^\top \Sigma_y^{-1} \widetilde{y}_i.\end{aligned}$$

Completing the square gives

$$z_i \mid (w_{i,S}, y_i, x_i) \sim \mathcal{N}(m_{i,S}, V_{i,S}).$$

Modalities-only posterior and posterior predictive for outcomes. Dropping the outcome term yields $z_i \mid (w_{i,S}, x_i) \sim \mathcal{N}(m_{i,S}^{(w)}, V_{i,S}^{(w)})$ with

$$V_{i,S}^{(w)-1} = \Sigma_0^{-1} + \sum_{m\in S} A_m^\top \Sigma_{w,m}^{-1} A_m, \qquad m_{i,S}^{(w)} = V_{i,S}^{(w)}\Big(\Sigma_0^{-1}\mu_0 + \sum_{m\in S} A_m^\top \Sigma_{w,m}^{-1}\widetilde{w}_{im}\Big).$$

Since $y_i \mid (z_i, x_i^s)$ is Gaussian and linear in z_i, integrating out $z_i \mid (w_{i,S}, x_i)$ yields

$$y_i \mid (w_{i,S}, x_i) \sim \mathcal{N}\Big(d(x_i^s) + C(x_i^s) m_{i,S}^{(w)},\ \Sigma_y + C(x_i^s) V_{i,S}^{(w)} C(x_i^s)^\top\Big).$$

Gaussian fusion and corrected product-of-experts exactness. For a single modality m, the unimodal posterior is $p_\vartheta(z_i \mid w_{im}, x_i) = \mathcal{N}(m_{im}, V_{im})$ with

$$V_{im}^{-1} = \Sigma_0^{-1} + A_m^\top \Sigma_{w,m}^{-1} A_m, \qquad V_{im}^{-1} m_{im} = \Sigma_0^{-1}\mu_0 + A_m^\top \Sigma_{w,m}^{-1}\widetilde{w}_{im}.$$

Define increments $\Delta_{im} := A_m^\top \Sigma_{w,m}^{-1} A_m$ and $h_{im} := A_m^\top \Sigma_{w,m}^{-1}\widetilde{w}_{im}$. Then the exact subset posterior (modalities only) satisfies

$$V_{i,S}^{(w)-1} = \Sigma_0^{-1} + \sum_{m\in S}\Delta_{im}, \qquad V_{i,S}^{(w)-1} m_{i,S}^{(w)} = \Sigma_0^{-1}\mu_0 + \sum_{m\in S} h_{im}.$$

Now consider the corrected product-of-experts rule from Eq. (19), specialized to Gaussian unimodal encoders that match the unimodal posteriors. Each unimodal posterior already contains

one copy of the prior factor $p_\psi(z_i \mid x_i^p)$; multiplying $|S|$ such terms duplicates the prior $|S|$ times. Dividing by $p_\psi(\cdot \mid x_i^p)^{|S|-1}$ removes the redundant prior factors, leaving exactly the prior times the product of the $|S|$ likelihood contributions. Since all factors are Gaussian in z_i, the resulting density is Gaussian and the precision algebra above shows it coincides with the exact subset posterior.

E.1 Proof of Proposition 1

Proof. Fix $\varepsilon > 0$. By the ELBO gap identity, for any $(d,x) \in \mathcal{D}_S \times \mathcal{X}$ and any $\lambda \in \Lambda$,

$$\log p_\vartheta(d \mid x) - \mathcal{L}(S;q_\lambda) = \mathrm{KL}\Big(q_\lambda \,\Big\|\, p_\vartheta(\cdot \mid d,x)\Big). \tag{23}$$

Thus it suffices to show there exists ϕ such that $\mathrm{KL}(q_{g_\phi(d,x)} \| p_\vartheta(\cdot \mid d,x)) \le \varepsilon$ uniformly over $(d,x) \in \mathcal{D}_S \times \mathcal{X}$.

Step 1: Construct a continuous near-optimal selector.

Define the best-in-family KL gap

$$\Delta_{\mathcal{Q}}(d,x) := \inf_{\lambda \in \Lambda} \mathrm{KL}\Big(q_\lambda \,\Big\|\, p_\vartheta(\cdot \mid d,x)\Big).$$

By condition (ii), $\Delta_{\mathcal{Q}}(d,x) = 0$ for all $(d,x) \in \mathcal{D}_S \times \mathcal{X}$.

By condition (iii), there exists a continuous function $\lambda^{\varepsilon/2} : \mathcal{D}_S \times \mathcal{X} \to \Lambda$ such that

$$\mathrm{KL}\Big(q_{\lambda^{\varepsilon/2}(d,x)} \,\Big\|\, p_\vartheta(\cdot \mid d,x)\Big) \le \Delta_{\mathcal{Q}}(d,x) + \frac{\varepsilon}{2} = \frac{\varepsilon}{2}$$

for all $(d,x) \in \mathcal{D}_S \times \mathcal{X}$.

Step 2: Establish uniform continuity.

Since $\mathcal{D}_S \times \mathcal{X}$ is compact and $\lambda^{\varepsilon/2}$ is continuous, the image

$$K := \lambda^{\varepsilon/2}(\mathcal{D}_S \times \mathcal{X}) \subseteq \Lambda$$

is compact. Define the set

$$K_\eta := \{\lambda \in \Lambda : \mathrm{dist}(\lambda, K) \le \eta\}$$

for $\eta > 0$. For η sufficiently small, K_η is compact (as a closed η-neighborhood of a compact set in $\mathbb{R}^p$).

By condition (iv), the map $(\lambda,d,x) \mapsto \mathrm{KL}(q_\lambda \| p_\vartheta(\cdot \mid d,x))$ is continuous. Since $K_1 \times \mathcal{D}_S \times \mathcal{X}$ is compact, this map is uniformly continuous on $K_1 \times \mathcal{D}_S \times \mathcal{X}$.

Therefore, there exists $\eta \in (0,1)$ such that for all $(d,x) \in \mathcal{D}_S \times \mathcal{X}$ and all $\lambda, \lambda' \in K_1$ with $\|\lambda - \lambda'\| < \eta$,

$$\Big|\mathrm{KL}\Big(q_\lambda \,\Big\|\, p_\vartheta(\cdot \mid d,x)\Big) - \mathrm{KL}\Big(q_{\lambda'} \,\Big\|\, p_\vartheta(\cdot \mid d,x)\Big)\Big| < \frac{\varepsilon}{2}. \tag{24}$$

Step 3: Approximate the selector with the encoder.

By condition (v), the encoder class can uniformly approximate the continuous selector $\lambda^{\varepsilon/2}$: $\mathcal{D}_S \times \mathcal{X} \to \Lambda$. Therefore there exists an encoder parameterization ϕ such that

$$\left\| g_\phi(d,x) - \lambda^{\varepsilon/2}(d,x) \right\| < \eta \quad \text{for all } (d,x) \in \mathcal{D}_S \times \mathcal{X}.$$

Step 4: Combine the bounds.

Fix any $(d,x) \in \mathcal{D}_S \times \mathcal{X}$. Let $\lambda' = \lambda^{\varepsilon/2}(d,x) \in K \subseteq K_1$ and $\lambda = g_\phi(d,x)$. By the choice of ϕ in Step 3, $\|\lambda - \lambda'\| < \eta < 1$, so $\lambda \in K_1$.

Applying the triangle inequality and the bounds from Steps 1 and 2:

$$\begin{aligned}
\mathrm{KL}\Big(q_{g_\phi(d,x)} \,\|\, p_\vartheta(\cdot \mid d,x)\Big) &\le \mathrm{KL}\Big(q_{\lambda^{\varepsilon/2}(d,x)} \,\|\, p_\vartheta(\cdot \mid d,x)\Big) \\
&\quad + \Big|\mathrm{KL}\Big(q_{g_\phi(d,x)} \,\|\, p_\vartheta(\cdot \mid d,x)\Big) - \mathrm{KL}\Big(q_{\lambda^{\varepsilon/2}(d,x)} \,\|\, p_\vartheta(\cdot \mid d,x)\Big)\Big| \\
&\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

Since (d,x) was arbitrary, this bound holds uniformly over $\mathcal{D}_S \times \mathcal{X}$. Combined with identity (23) and the nonnegativity of KL divergence, we obtain

$$0 \le \log p_\vartheta(d \mid x) - \mathcal{L}(S; q_{g_\phi(d,x)}) \le \varepsilon$$

for all $(d,x) \in \mathcal{D}_S \times \mathcal{X}$. □

F Proofs for the extended linear–Gaussian case

We prove Theorems 1 and 2 using standard arguments for M-estimators applied to the augmented i.i.d. data $U_i = (d_{i,S}, x_i, \varepsilon_i)$.

F.1 Population objective and profiling identity

Let $q_\phi(z \mid d_S, x)$ be any candidate encoder distribution and denote the exact posterior by $p_\vartheta(z \mid d_S, x)$. The ELBO identity implies (pointwise in (d_S, x))

$$\log p_\vartheta(d_S \mid x) = \mathcal{L}(S; q_\phi) + \mathrm{KL}\big(q_\phi(\cdot \mid d_S, x) \,\|\, p_\vartheta(\cdot \mid d_S, x)\big).$$

Taking expectations over the data distribution yields the decomposition of the population objective

$$\mathcal{J}(\vartheta, \phi) = \mathbb{E}\left[\log p_\vartheta(d_S \mid x)\right] - \mathbb{E}\left[\mathrm{KL}\big(q_\phi(\cdot \mid d_S, x) \,\|\, p_\vartheta(\cdot \mid d_S, x)\big)\right].$$

Under Assumption 1(v), for each ϑ near ϑ_0 there exists $\phi^*(\vartheta)$ with zero KL term almost surely, so

$$\sup_\phi \mathcal{J}(\vartheta, \phi) = \mathbb{E}\left[\log p_\vartheta(d_S \mid x)\right].$$

This is the precise sense in which (under exact inference / posterior realizability and when amortization does not bind) profiling out ϕ yields an objective over ϑ that shares maximizers with maximum likelihood.

F.2 Proof of Theorem 1

Proof. Write $\eta = (\vartheta, \phi)$ and $\widehat{\mathcal{J}}_n(\eta) = \frac{1}{n}\sum_{i=1}^n m(\eta; U_i)$ as in Eq. (17). By Assumption 1(i)–(ii) and standard uniform law of large numbers conditions (continuity in η and an integrable envelope), we have

$$\sup_{\eta \in \mathcal{H}} \left| \widehat{\mathcal{J}}_n(\eta) - \mathcal{J}(\eta) \right| \xrightarrow{p} 0.$$

Next, by the profiling identity in Appendix F.1,

$$\sup_{\phi} \mathcal{J}(\vartheta, \phi) = \mathbb{E}\left[\log p_\vartheta(d_S \mid x)\right]$$

in a neighborhood of ϑ_0. Under correct specification and identification up to rotations (Assumption 1(iii)–(iv)), the population log-likelihood $\mathbb{E}[\log p_\vartheta(d_S \mid x)]$ is maximized precisely on the equivalence class $[\vartheta_0]$.

Finally, by the argmax theorem for M-estimators on compact parameter spaces, any maximizer $\widehat{\eta}_n$ of $\widehat{\mathcal{J}}_n$ converges in probability to the set of population maximizers

$$\mathcal{M} = \{(\vartheta, \phi) : \vartheta \in [\vartheta_0],\ q_\phi(\cdot \mid d_S, x) = p_\vartheta(\cdot \mid d_S, x) \text{ a.s.}\},$$

which gives $\operatorname{dist}(\widehat{\eta}_n, \mathcal{M}) \to_p 0$. □

F.3 Proof of Theorem 2

Proof. This is a standard smooth M-estimation CLT applied to the augmented i.i.d. data $U_i = (d_{i,S}, x_i, \varepsilon_i)$. Under Assumption 1(vi), a Taylor expansion of the first-order condition $\nabla_\eta \widehat{\mathcal{J}}_n(\widehat{\eta}_n) = 0$ around an identified representative η_0 yields

$$0 = \nabla_\eta \widehat{\mathcal{J}}_n(\eta_0) + \left(\nabla^2_{\eta\eta} \widehat{\mathcal{J}}_n(\tilde{\eta}_n)\right)(\widehat{\eta}_n - \eta_0),$$

for some intermediate $\tilde{\eta}_n$ between $\widehat{\eta}_n$ and η_0. Multiplying by $\sqrt{n}$ and using (i) a multivariate CLT for $\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\eta m(\eta_0; U_i)$ with variance B, and (ii) convergence in probability of $\nabla^2_{\eta\eta} \widehat{\mathcal{J}}_n(\tilde{\eta}_n)$ to $-A$, gives

$$\sqrt{n}(\widehat{\eta}_n - \eta_0) \Rightarrow \mathcal{N}(0, A^{-1} B A^{-1}).$$

□