Easy Variational Inference for Categorical Models via an Independent Binary Approximation

Michael T. Wojnowicz [1 2] **Shuchin Aeron** [1 3] **Eric L. Miller** [1 3] **Michael C. Hughes** [1 4]

Abstract

We pursue tractable Bayesian analysis of generalized linear models (GLMs) for categorical data. GLMs have been difficult to scale to more than a few dozen categories due to non-conjugacy or strong posterior dependencies when using conjugate auxiliary variable methods. We define a new class of GLMs for categorical data called *categorical-from-binary* (CB) models. Each CB model has a likelihood that is bounded by the product of binary likelihoods, suggesting a natural posterior approximation. This approximation makes inference straightforward and fast; using well-known auxiliary variables for probit or logistic regression, the product of binary models admits conjugate closed-form variational inference that is embarrassingly parallel across categories and invariant to category ordering. Moreover, an independent binary model simultaneously approximates *multiple* CB models. Bayesian model averaging over these can improve the quality of the approximation for any given dataset. We show that our approach scales to thousands of categories, outperforming posterior estimation competitors like Automatic Differentiation Variational Inference (ADVI) and No U-Turn Sampling (NUTS) in the time required to achieve fixed prediction quality.

1 Introduction

We consider the problem of modeling categorical data informed by covariates using the machinery of generalized linear models (GLMs). Because our intended big data applications may involve rare events or little data for some quantities of interest, we take a Bayesian approach in order to estimate *distributions* over unknown parameters given available data, and then average over these distributions when making predictions. While many generalized linear models for categorical data have been proposed, Bayesian analysis of these models remains difficult with substantial active research due to the need for methods that are simultaneously accurate, tractable, and scalable.

The most common modeling choice for categorical data is multi-class logistic regression, which uses a softmax (a.k.a. multi-logit) function to produce category probabilities. The softmax likelihood is not conjugate to any standard prior over weight parameters (such as Gaussian), so estimating posteriors over weights requires expensive sampling methods (Hoffman et al., 2014) or non-conjugate variational optimization methods (Wang & Blei, 2013; Braun & McAuliffe, 2010; Kucukelbir et al., 2017). Recent auxiliary variable methods (Polson et al., 2013) have yielded conjugate conditionals amenable to Gibbs sampling, but closed-form variational updates for multiple categories require *stick-breaking* (Linderman et al., 2015). Stick-breaking imposes an asymmetric order over categories, yet in many cases it is unnatural to view category selection as a sequential process. In practice, this asymmetry complicates prior specification and inference quality (Zhang & Zhou, 2017).

An alternative model is multi-class probit regression, whose link function is the cumulative distribution function of the Normal distribution. The probit admits conjugate inference under a well-known auxiliary variable representation (Albert & Chib, 1993; Held & Holmes, 2006). However, multi-class probit models encode strong posterior dependence among entries of the auxiliary parameter vectors. This dependence requires one-entry-at-a-time sampling instead of joint sampling (Johndrow et al., 2013), yielding poor mixing performance as the number of categories grows. Furthermore, implementations often require picking a "base category"; this choice can impact the practical results of inference (Burgette et al., 2021). Finally, the multinomial probit lacks closed-form category probabilities (Johndrow et al., 2013), which has prevented adoption within more complicated models (Holsclaw et al., 2017).

Motivated by difficulties that arise from these previous efforts (summarized in Table 1), we present a new class of categorical models – *categorical-from-binary* (CB) models[1] – whose defining feature is that each one's likelihood

[1]Tufts University, Medford, MA, USA [2]Data Intensive Studies Center [3]Dept. of Electrical and Computer Engineering [4]Dept. of Computer Science. Email: <michael.wojnowicz@tufts.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Code: github.com/tufts-ml/categorical-from-binary

Table 1: Assessment of categorical regression models in terms of the presence (✓) or absence (✗) of desirable features for fast, scalable Bayesian inference. *Rows:* PGA refers to Pòlya-Gamma augmentation (Polson et al., 2013). SB-Softmax refers to softmax regression with a stick-breaking link function (Linderman et al., 2015). MNP+ACA stands for multinomial probit with Albert & Chib (1993) augmentation. IB refers to our proposed independent binary regression (Sec. 2.2). The first two rows are categorical models, the next three rows are categorical models with augmentation, and the last two rows are not categorical models, but in this paper we show how (and justify why) they can be used for approximate inference. See Sec. F.1 for an extended version of this table caption.

Model	**Inference Feature**						
	Invariance to category ordering	Latent linear regression	Auxiliary variable independence	Closed-form likelihood	Conditional conjugacy	Closed-form variational inference	Embarassingly parallel across categories
Softmax	✓	✗	✓	✓	✗	✗	✗
MNP	✓	✗	✓	✗	✗	✗	✗
Softmax+PGA	✓	✓	✓	✓	✓	✗	✗
SB-Softmax+PGA	✗	✓	✓	✓	✓	✓	✗
MNP+ACA	✓	✓	✗	✗	✓	✓	✗
IB-Probit+ACA	✓	✓	✓	✓	✓	✓	✓
IB-Logit+PGA	✓	✓	✓	✓	✓	✓	✓

can be lower-bounded by the likelihood of an independent binary model. To perform approximate posterior estimation for such models, we fit the independent binary model via coordinate-ascent variational methods, taking advantage of well-known closed-form updates for binary logit or probit models. This approach is scalable to thousands of categories, even more so because it is *embarrassingly parallel* across categories, meaning we can fit a separate model for each category with no inter-worker communication overhead (Foster, 1995). Even without parallelization, we demonstrate heldout predictions of comparable prediction quality to other categorical GLMs in far less time (see Fig. 3), with competitive likelihoods only slightly below the expensive gold standards. Our accurate predictions are possible via a *Bayesian model average* (BMA) over members of our CB model class which we can deploy cheaply using only one posterior fit for the surrogate model. Our experiments reveal that our proposed methods offer a promising first-line approach for fast Bayesian analysis of big categorical data, especially when the number of categories is large.

1.1 Problem formulation

Consider a given training set of N paired observations, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where each observation (indexed by i) consists of $\boldsymbol{x}_i \in \mathbb{R}^M$, a (fixed) vector of covariates, and integer $y_i \in \{1, ..., K\}$, indicating which of the K categories i belongs to. We treat y_i as a random variable generated as:

$$y_i \sim \text{Cat}(\boldsymbol{s}_i), \quad \boldsymbol{s}_i = (s_{i1}, \dots s_{iK})^T \in \Delta_{K-1}, \quad \text{(1.1.1a)}$$

$$\boldsymbol{s}_i = f(\boldsymbol{\eta}_i), \quad \boldsymbol{\eta}_i = \boldsymbol{B}^T \boldsymbol{x}_i. \quad \text{(1.1.1b)}$$

Here, $\boldsymbol{B} \in \mathbb{R}^{M \times K}$ are unobserved regression weights, whose matrix-vector product with covariates $\boldsymbol{x}_i$ yields the so-called linear predictor $\boldsymbol{\eta}_i$. The function $f : \mathbb{R}^K \to \Delta_{K-1}$ maps the real-valued vector $\boldsymbol{\eta}_i$ to a vector $\boldsymbol{s}_i$ of K non-negative values that sum to one. We refer to this model as a *categorical regression* or a generalized linear model (GLM) for categorical data. Note that f need not be invertible, so the model need not be identifiable (i.e., there may exist $\boldsymbol{B}_1 \neq \boldsymbol{B}_2$ which yield identical distributions over y_i).

We wish to pursue Bayesian inference, treating the parameter $\boldsymbol{B}$ as a random variable with prior $\pi(\boldsymbol{B})$. We use a Gaussian prior in practice. Our primary interest is the *prediction task*: given N training pairs $(\boldsymbol{x}_i, y_i)$ and a new covariate vector $\boldsymbol{x}_*$, we wish to make probabilistic prediction of the new category label y_* via the posterior predictive $p(y_*|\{y_i\}_{i=1}^N) = \int p(y_*|\boldsymbol{B})p(\boldsymbol{B}|\{y_i\}_{i=1}^N)d\boldsymbol{B}$. This prediction requires the completion of a *posterior estimation* task: given a fixed training set of size N, estimate $p(\boldsymbol{B}|\{y_i\}_{i=1}^N)$. To keep the formal statements of both tasks simple, we treat covariates as fixed knowns and suppress conditioning on $\boldsymbol{x}_i$ in notation. We stress that our focus is on the posterior predictive, as category outcomes y_* are relevant to applications while the weights $\boldsymbol{B}$ are intermediate quantities whose non-identifiability can make assessment challenging; large differences in parameter space may not imply notable changes in prediction quality.

Contributions. Our contribution is to define a class of categorical models (choices of the function f) that we call *categorical-from-binary* models. Using this class, we show that a well-justified approximation is possible such that posterior estimation enjoys all the beneficial properties in Table 1. To our knowledge, out of all previous models

listed in Table 1, only our approach yields tractable category probabilities, provides scalable yet closed-form variational optimization, is invariant to category order, and can be integrated into more complex graphical models. We further provide a prediction method that averages across models to obtain accurate categorical predictions from our approximate posterior.

2 Models

2.1 Overview

Bayesian inference for *categorical* regressions (Eq. (1.1.1)) is difficult, in the sense that no current approach has all the features given in Table 1. In contrast, Bayesian inference for *binary* regression is far more straightforward. Using Pòlya-Gamma augmentation (Polson et al., 2013) for logistic regression, or the Normal augmentation of Albert & Chib (1993) for probit regression, one may obtain conditionally conjugate models and closed-form variational inference (Durante & Rigon, 2019; Consonni & Marin, 2007; Armagan & Zaretzki, 2011; Fasano et al., 2019). These properties extend to regression models for K-bit binary vectors that treat each bit independently (see Sec. 2.2). In fact, if each bit were to indicate the presence of a particular category, then every desirable inferential feature listed in Table 1 would be present. We would like to exploit this collection of features for *categorical* modeling. The problem is that independent binary regression allows for multiple non-zero bits while categorical models require exactly one non-zero bit. To address this problem, we construct categorical models around independent binary models (Sec. 2.3), which enables the efficient posterior estimation (Sec. 3).

2.2 A model for independent binary vectors

Consider a general univariate binary regression likelihood (Albert & Chib, 1993) of the form

$$\breve{y}_i \mid \breve{\boldsymbol{\beta}} \overset{\text{ind}}{\sim} \text{Bernoulli}\big(H(\breve{\eta}_i)\big), \quad i = 1, ..., n \tag{2.2.1}$$

where $\breve{y}_i \in \{0, 1\}$ are binary response random variables, the linear predictor $\breve{\eta}_i = \boldsymbol{x}_i^T \breve{\boldsymbol{\beta}}$ is formed from known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\breve{\boldsymbol{\beta}} \in \mathbb{R}^M$, and H is an arbitrary cdf that is referred to as an inverse link function. Logistic regression sets H to be the standard logistic cdf and probit regression sets H to the standard Gaussian cdf. We use the breve notation to distinguish random variables here from those in later categorical models. For additional intuition about Eq. (2.2.1), see Sec. A.

Now let us consider modeling binary *vectors*: $\widehat{\boldsymbol{y}}_i = (\widehat{y}_{i1}, ..., \widehat{y}_{iK}) \in \{0, 1\}^K, i = 1, 2, \ldots, N$. Crucially, each $\widehat{\boldsymbol{y}}_i$ is a K-bit vector, and *not* a one-hot vector: any number of entries could be 1 or 0. Taking the product of binary regression likelihoods of the form in Eq. (2.2.1), we obtain a model which we call *independent binary* (IB) *regression*,

$$\widehat{y}_{ik} \mid \widehat{\boldsymbol{\beta}}_k \overset{\text{ind}}{\sim} \text{Bernoulli}\big(H(\widehat{\eta}_{ik})\big) \tag{2.2.2}$$

independently across each $k = 1, 2, \ldots, K$, with each linear predictor $\widehat{\eta}_{ik} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k$ formed from known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\widehat{\boldsymbol{\beta}}_k \in \mathbb{R}^M$. The likelihood for an observation under a K-bit IB model is

$$p_{\text{IB}}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{B}}) = \prod_{k=1}^{K} H(\widehat{\eta}_{ik})^{\widehat{y}_{ik}} \left(1 - H(\widehat{\eta}_{ik})\right)^{1-\widehat{y}_{ik}}, \tag{2.2.3}$$

where $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1, ..., \widehat{\boldsymbol{\beta}}_K) \in \mathbb{R}^{M \times K}$ is a matrix of weights for each combination of covariate and category. IB is a *class* of models, each member defined by a chosen H. When H is the standard logistic or standard Gaussian cdf, we respectively obtain the IB-Logit or IB-Probit models.

2.3 Categorical-from-binary models

Suppose now that we are interested in regression models for categorical (one-of-K) data, i.e. $y_i \sim \text{Cat}(s_{i1}, \ldots s_{iK})$ where $y_i \in \{1, ..., K\}$ and $\boldsymbol{s}_i \in \Delta_{K-1}$. We restrict our focus to categorical models which are related to IB models (Eq. (2.2.2)) in the following manner:

Definition 2.3.1. A **categorical-from-binary** (CB) model is a GLM for categorical data $y_i \in \{1, .., K\}$ which always assigns a higher likelihood to a category k than an IB model does to the corresponding one-hot vector. That is, CB models obey the likelihood bound

$$p_{\text{CB}}(y_i \mid \boldsymbol{B}) > p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_{y_i} \mid \widehat{\boldsymbol{B}} = \boldsymbol{B}) \tag{2.3.1}$$

for all observations $y_i \in \{1, ..., K\}$, covariates $\boldsymbol{x}_i \in \mathbb{R}^M$, and weights $\boldsymbol{B} \in \mathbb{R}^{M \times K}$, and where $\boldsymbol{e}_{y_i}$ is the one-hot indicator vector with value of 1 only at entry y_i. △

To construct a CB model, we must construct a function f for the relation $p_{\text{CB}}(y_i \mid \boldsymbol{B}) = f(\boldsymbol{\eta}_i)$, where $\boldsymbol{\eta}_i = \boldsymbol{B}^T \boldsymbol{x}_i$, such that the bound in Eq. (2.3.1) is satisfied. We begin by choosing a cdf H (e.g. standard Gaussian or standard Logistic) to specify a concrete IB model (e.g. IB-Probit or IB-Logit). We refer to the chosen IB-model as the **base** of a CB model. We then define $h : \mathbb{R}^K \to \mathbb{R}^K$ such that $h(\boldsymbol{\eta}_i) = \big(H(\eta_{i1}), \ldots, H(\eta_{iK})\big)^T$. CB models construct f via the composition $f = g \circ h$ for some function g defined below. This composition means that CB category probabilities are determined by the vector output of h, whose entries define the probabilities of "success" at each of the K bits of the IB model: $H(\eta_{ik}) = p_{\text{IB}}(\widehat{y}_{ik} = 1 \mid \boldsymbol{\beta}_k)$ for all k.

2.4 Concrete categorical-from-binary likelihoods

After selecting a specific cdf H, fully specifying a concrete CB model for categorical data requires identifying the transformation g which maps the IB probabilities of success $h(\boldsymbol{\eta}_i)$ into the simplex Δ_{K-1} in a way that satisfies the bound in Eq. (2.3.1). We now provide two such specifications. First, the *marginalization* construction assumes the probability of the kth category is proportional to the

probability of success in the k-th bit of the IB model. Second, the *conditioning* construction requires the IB model to assign non-zero probability only to valid one-hot vectors, and then assumes that the probability of the kth category equals the probability of its one-hot representation.

CB **construction via marginalization.** A *categorical-from-binary-via-marginalization* (CBM) model produces category probabilities by normalizing the marginal probabilities of success $\{H(\eta_{ik})\}_{k=1}^K$ from an IB model:

$$p_{\text{CBM}}(y_i = k \mid \boldsymbol{B}) = \frac{H(\eta_{ik})}{\sum_{\ell=1}^K H(\eta_{i\ell})}. \tag{2.4.1}$$

for all categories $k \in \{1, \ldots, K\}$.

CB **construction via conditioning.** A *categorical-from-binary-via-conditioning* (CBC) model produces category probabilities from an IB model by conditioning on the event that the vector has exactly one positive entry:

$$p_{\text{CBC}}(y_i{=}k \mid \boldsymbol{B}) = \frac{H(\eta_{ik}) \prod_{j \neq k}(1 - H(\eta_{ij}))}{\sum_{\ell=1}^{K} H(\eta_{i\ell}) \prod_{j \neq \ell}(1 - H(\eta_{ij}))} \tag{2.4.2}$$

for categories $k = 1, ..., K$. A CBC model is an IB model truncated to the space of one-hot vectors. A CBC model can also be expressed as a *normalized odds* model (see Sec. B.2).

Proposition 2.4.1. CBC *and* CBM *models are categorical-from-binary (*CB*) models satisfying Definition 2.3.1.*

Proof. Deferred to Appendix Sec. B.3 to save space. □

Example 2.4.1. To illustrate the strategies taken by the CBM and CBC models in forming a categorical regression from an IB regression, contrast how they assign probability to the first of $K = 3$ categories.

$$p_{\text{CBM}}(y = 1 \mid \boldsymbol{B}) \propto p_{\text{IB}}\Big(\widehat{\boldsymbol{y}} \in \{(1,0,0), (1,1,0), (1,0,1), (1,1,1)\} \mid \boldsymbol{B}\Big)$$

$$p_{\text{CBC}}(y = 1 \mid \boldsymbol{B}) = p_{\text{IB}}\Big(\widehat{\boldsymbol{y}} = (1,0,0) \,\Big|\, \widehat{\boldsymbol{y}} \in \{(1,0,0), (0,1,0), (0,0,1)\}, \boldsymbol{B}\Big)$$

The distinction can be understood through a voting metaphor. The conditioning strategy of CBC models forces the K independent binary models to agree on a single "vote", the marginalization strategy of CBM models allows multiple votes across the K independent bits. △

From model classes to models. Both CBC and CBM are model *classes*, generating different models as H varies. For instance, by taking H to be the standard Gaussian cdf we can generate the CBC-Probit and CBM-Probit models (for which IB-Probit is the base), and by taking H to be the standard Logistic cdf, we can generate the CBC-Logit and CBM-Logit models (for which IB-Logit is the base).

2.5 Related work

Diagonal orthant models. Our work is inspired by the *diagonal orthant* (DO) models proposed by Johndrow et al. (2013). What we call the CBC likelihood is equivalent to the marginal likelihood of the DO model (integrating away auxiliary variables). Johndrow et al. further proposed using independent binary regressions (as we do) to perform scalable Bayesian computation for categorical data. Johndrow et al. argued for IB approximation based on a claimed identification equivalence of point estimated weights between IB and DO models.

In a recent non-archival workshop paper (Wojnowicz et al., 2021), we clarified that IB should be viewed as a separate, surrogate model (see also Sec. B.4). This paper extends that early line of work, offering a more coherent view of surrogate bounds, expanding to include many possible cdfs (not just probit) for IB approximations, and introducing our BMA approach to effective predictions. We also simplify inference, as neither the auxiliary variables in Johndrow et al.'s DO model (nor the auxiliary variables in (Wojnowicz et al., 2021)'s SDO model) are needed to relate IB models to categorical models.

In summary, building on the IB inference strategy first suggested by Johndrow et al., we contribute the following advances: (1) We clarify that doing inference on a relevant categorical model via an IB model requires an *approximation*. (2) We *justify* this approximation via surrogate likelihood bounds. (3) We expand the class of categorical models suitable for IB approximation, showing that *both* CBC and CBM models should be included to obtain high-quality predictions (see Sec. 5.1). (4) We focus on optimization approaches to posterior estimation, which may be more scalable than Johndrow et al.'s Gibbs sampling.

Three-step augmentation. Galy-Fajou et al. (2020) pursue conjugate inference for multi-class Gaussian process classification using a categorical likelihood that is identical to CBM-Logit. They obtain exact inference (without any IB approximation) using a model augmentation strategy with three hierarchical levels. This strategy also applies without Gaussian processes. However, our approach remains conceptually simpler and appears more scalable due to parallelization and use of the probit link. Future work is needed to assess the tradeoffs in practice.[2]

One-vs-rest classification. At a high-level, our IB approximation is similar to a common generic heuristic for building multi-class classifiers known as a *one-versus-rest* (or *one-versus-all*) ensemble. One-vs-rest schemes fit K separate binary classifiers to distinguish each class from all others, and then make a one-of-K prediction by taking the class corresponding to the classifier with largest predicted

[2] We discovered this paper just before publication.

score or probability. Empirically, one-vs-rest schemes can deliver accuracies on par with one-of-K classifiers for non-linear kernel methods (Rifkin & Klautau, 2004). Due to simplicity and computational speed, widely-used software packages support this scheme (Pedregosa et al., 2011) and efforts to classify 10,000 image classes have called one-vs-rest "the only affordable option" (Deng et al., 2010). However, to our knowledge there has not yet been statistical justification for one-vs-rest schemes in terms of a true multi-class likelihood, leading to concerns about coherency (Murphy, 2012). Our framework provides *probabilistic* one-vs-rest approximations of categorical models.

3 Posterior estimation

We now develop our methodology for approximating a CB model's posterior over weights, $p(\boldsymbol{B}|\{y_i\}_{i=1}^N)$. The key insight is this: we can provably optimize a lower bound on the likelihood of a CB model by instead performing traditional variational inference for the IB model. First, we establish that after integrating away the weights $\boldsymbol{B}$, the marginal likelihood of a CB model is lower-bounded by the marginal likelihood of an IB model. Second, we argue that a mean-field variational posterior $q(\boldsymbol{B})$ estimated to approximate the IB model is also a suitable approximation for a CB model. This suggests a straightforward coordinate ascent variational inference algorithm (IB-CAVI), which uses the efficient conjugate updates for logit or probit *binary* models discussed in Sec. 2.1.

3.1 Marginal likelihood bounds

For any dataset $\boldsymbol{y} = \{y_i\}_{i=1}^N$, any CB likelihood and any choice of prior with density $\pi(\boldsymbol{B})$, we obtain

$$p_{\text{CB}}(\boldsymbol{y}) = \int \pi(\boldsymbol{B}) \prod_{i=1}^N p_{\text{CB}}(y_i \mid \boldsymbol{B})\, d\,\boldsymbol{B} \tag{3.1.1a}$$

$$> \int \pi(\boldsymbol{B}) \prod_{i=1}^N p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_{y_i} \mid \boldsymbol{B})\, d\,\boldsymbol{B} \tag{3.1.1b}$$

$$= p_{\text{IB}}(\widehat{\boldsymbol{Y}} = \boldsymbol{E}(\boldsymbol{y})), \tag{3.1.1c}$$

which follows from the likelihood bound relating CB to IB (Eq. (2.3.1)) and monotonicity of the integral. Here, $\widehat{\boldsymbol{Y}} = (\widehat{\boldsymbol{y}}_i)_{i=1}^N$, and $\boldsymbol{E}(\boldsymbol{y})$ represents a one-hot representation of the categorical training data $\boldsymbol{y}$, stacking all one-hot vectors $\{\boldsymbol{e}_{y_i}\}_{i=1}^N$.

3.2 Variational surrogate bounds

Variational inference (Wainwright et al., 2008; Blei et al., 2017) deterministically approximates a posterior distribution by finding the member $Q \in \mathcal{Q}$ of a tractable family of distributions which maximizes a lower bound on the logarithm of the *evidence* (the marginal likelihood of the data). This lower bound is known as the evidence lower bound or "ELBO".

For categorical-from-binary (CB) models, the evidence of interest is $p_{\text{CB}}(\boldsymbol{y}) = \int p_{\text{CB}}(\boldsymbol{y}|\boldsymbol{B})\, \pi(\boldsymbol{B})\, d\boldsymbol{B}$, where π denotes the prior density on $\boldsymbol{B}$. This quantity is intractable for both CBC and CBM models (as defined in Sec. 2.4) because they lack a conjugate prior. For instance, a Gaussian prior is not conjugate, since the logarithm of the joint density $p_c(y_i \mid \boldsymbol{B})\pi(\boldsymbol{B})$ does not yield a quadratic in $\boldsymbol{B}$, where c is in $\{\text{CBC}, \text{CBM}\}$ and π is a Gaussian density.

Bayesian inference for a model with an intractable marginal likelihood can sometimes be provided through a conventional variational approach. If we select Q as any distribution over $\boldsymbol{B} \in \mathbb{R}^{M\times K}$, and let $q(\cdot)$ be the density of Q, then the traditional lower bound of the log of the evidence, which we denote $\texttt{ELBO}_{\text{CB}} \leq \log p_{\text{CB}}(\boldsymbol{y})$, follows from Jensen's inequality:

$$\texttt{ELBO}_{\text{CB}}(q) = \mathbb{E}_q[\log p_{\text{CB}}(\boldsymbol{y} \mid \boldsymbol{B})] - D_{\text{KL}}(q(\boldsymbol{B}) \parallel \pi(\boldsymbol{B})), \tag{3.2.1}$$

where D_{KL} is the Kullback-Leibler divergence. Unfortunately, this lower bound is still intractable to compute. The energy term $\mathbb{E}_q[\log p_{\text{CB}}(\boldsymbol{y} \mid \boldsymbol{B})]$ contains N expectations that lack closed-form expression (expected log sums of K nonlinear quantities), due to the normalizing constants of the categorical models (Eqs. (2.4.1) and (B.2.2)). While Monte Carlo approximations to this integral are possible that can enable gradient-based learning of $q(\boldsymbol{B})$, given a fixed computational budget the quality of these approximations becomes increasingly suspect in high dimensions, such as when the number of categories grows.

Instead, we define a surrogate objective $\mathcal{L}_{\text{CB}}(q)$ that lower bounds the log marginal likelihood for any CB model:

$$\log p_{\text{CB}}(\boldsymbol{y}) \overset{(3.1.1)}{>} \log p_{\text{IB}}(\widehat{\boldsymbol{Y}} = \boldsymbol{E}(\boldsymbol{y})) \tag{3.2.2a}$$

$$\geq \texttt{ELBO}_{\text{IB}}(q; \widehat{\boldsymbol{Y}} = \boldsymbol{E}(\boldsymbol{y})) := \mathcal{L}_{\text{CB}}(q). \tag{3.2.2b}$$

We call this a *surrogate lower bound* because there are two bounds at work: the bound relating CB to IB in Eq. (3.2.2a) and the traditional ELBO (via Jensen's inequality) in Eq. (3.2.2b). (Recall from Eq. (3.1.1) that the the former bound requires that the CB model and its IB base have the *same* prior density π over weights.) This yields a surrogate objective $\mathcal{L}_{\text{CB}}(q)$ which is exactly the traditional ELBO applied to the IB model. As justified by this surrogate, we can solve our Bayesian inference problem for *categorical* regression by applying well-known variational inference scheme for *binary* regression on a one-hot transformation of the categorical data.

3.3 Procedure for posterior estimation

We now outline scalable procedures for closed-form coordinate ascent variational inference (CAVI) that will estimate an optimal approximate posterior $q^*(\boldsymbol{B})$ under the surrogate objective $\texttt{ELBO}_{\text{IB}}$. Especially in high-dimensional settings, these procedures are far more scalable than the difficult task of directly optimizing the truly-categorical model (via the objective $\texttt{ELBO}_{\text{CB}}$).

Closed-form CAVI procedures for univariate binary re-

gression (Eq. (2.2.1)) are well-known when the prior on weights $\breve{\boldsymbol{\beta}}$ is Gaussian and the function H corresponds to either logit or probit regression, as reviewed in Sec. 2.1. Both involve augmenting observed binary data $\breve{\boldsymbol{y}} \in \{0,1\}^N$ with auxiliary variables $\breve{\boldsymbol{z}} \in \mathbb{R}^N$ such that the augmented model is conditionally conjugate while the original model is preserved through marginalization. In particular, we can obtain a conditionally conjugate model by either constructing $\breve{\boldsymbol{z}}$ via truncated normal augmentation for probit regression (Albert & Chib, 1993) or via Pòlya-Gamma augmentation for logistic regression (Polson et al., 2013). Extrapolating from this univariate binary model to the K independent bits of the IB model (Eq. (2.2.3)), we immediately obtain conditional conjugacy by augmentation with a matrix $\widehat{\boldsymbol{Z}} \in \mathbb{R}^{N\times K}$ whose kth column $\widehat{\boldsymbol{z}}_k$ uses the relevant univariate augmentation strategy.

Our variational approximation of the *augmented* K-bit IB model assumes a mean-field factorization: $q(\widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{B}}) = q(\widehat{\boldsymbol{Z}})q(\widehat{\boldsymbol{B}})$. Under this choice, deriving a coordinate ascent algorithm to find the q that optimizes $\mathcal{L}_{\text{CB}}(q)$ in Eq. (3.2.2) follows the standard variational recipe (Blei et al., 2017). First, because both prior and likelihood are independent across categories k, our mean-field posterior also simplifies as independent across categories without further approximation: $q(\widehat{\boldsymbol{Z}}, \widehat{\boldsymbol{B}}) = \prod_{k=1}^{K} q(\widehat{\boldsymbol{z}}_k)q(\widehat{\boldsymbol{\beta}}_k)$. From there, one exploits known applications of CAVI to univariate binary models specific to the chosen link function, either logit (Durante & Rigon, 2019) or probit (Consonni & Marin, 2007; Armagan & Zaretzki, 2011; Fasano et al., 2019). Procedurally, from a suitable initial value of q, each factor of q is updated to maximize $\texttt{ELBO}_{\text{IB}}$ while holding others fixed, using closed-form updates arising from conditional conjugacy. For concrete realizations of the required updates for a K-bit IB model, see Sec. D for the probit link and Sec. E for the logit. Since this posterior estimation procedure operates by doing CAVI for a surrogate IB model, we call it *IB-CAVI (independent binary coordinate ascent variational inference)*.

After iterating updates to each factor until convergence, the resulting variational density q^* over $\widehat{\boldsymbol{B}}$ is a local maximum of the $\texttt{ELBO}_{\text{IB}}$ (Ormerod & Wand, 2010). By Eq. (3.2.2), $q^*(\widehat{\boldsymbol{B}})$ is therefore a local maximum of a surrogate bound on the CB model, and thus we can treat $q^*(\widehat{\boldsymbol{B}})$ as an approximation to the ideal (intractable) posterior $p_{\text{CB}}(\boldsymbol{B}|\boldsymbol{y})$ of the categorical model.

Our IB-CAVI procedure is not specific to a particular CB model. One optimization run can produce a posterior q^* suitable for *multiple* CB target models, as long as the IB model is a base. For example, performing CAVI for the IB-Probit model provides a q^* suitable for *any* CB-Probit model (CBM-Probit, CBC-Probit, etc.).

Runtime cost of IB-CAVI. The per-iteration runtime cost for logit models is $O(M^3K + NM^2K)$ (see Alg. 3 and Sec. E.4), where K is the number of categories, M is the number of features, and N is the number of training examples. For probit models, the per-iteration runtime drops to $O(NMK)$, with further reductions under sparsity (Alg. 2 and Sec. D.4). When the Gaussian prior $\pi(\boldsymbol{B})$ is chosen to be independent across category-specific weights, under either link function our IB-CAVI approach is *embarrassingly parallel* across categories. This makes our IB-CAVI approach particularly suitable for data with hundreds or thousands of categories. For example, to fit the IB-Probit in parallel, each worker solves a single category's binary regression problem to convergence at cost $O(NM)$ per iteration.

4 Prediction via Bayesian Model Averaging

Given a posterior over weights $q^*(\boldsymbol{B})$ via the IB-CAVI procedure from Sec. 3, how can we make useful *predictions* of the category labels $y_* \in \{1 \dots K\}$ for new observations with covariates $\boldsymbol{x}_*$? Clearly we must employ a truly-categorical CB likelihood to obtain valid predictions, as the IB likelihood can produce any K-bit vector, not just a 1-of-K choice. However, empirical investigations in Sec. B.5.1 (see esp. Fig. B.1), with further results in Fig. 1, suggest that there are substantial dataset-specific tradeoffs in approximation quality (IB can approximate CBC better than CBM on some data, and vice versa on other data) and goodness-of-fit. Needing to select a specific CB likelihood (CBC or CBM) in advance for a dataset would be challenging.

To avoid this problem, we exploit the fact that our IB-CAVI procedure produces a posterior q^* suitable for *multiple* CB likelihoods. Thus, to make predictions we perform a *Bayesian model average* over all applicable CB likelihoods. We find this significantly improves prediction quality at *no additional training cost*.

We model the problem with two random variables. First, let $c \in \{\text{CBM}, \text{CBC}\}$ indicate the selected model, with given prior probabilities $p(c) = \pi_c \in (0,1)$ such that $\sum_c \pi_c = 1$. We recommend a uniform setting: $\pi_{\text{CBM}} = \pi_{\text{CBC}} = 0.5$.[3] Second, we have the predicted quantity of interest Δ (such as future category label y_*), for which $p_c(\Delta|\boldsymbol{B})$ is known. Following Madigan et al. (1996), our BMA prediction procedure forms the posterior predictive for Δ given training data $\boldsymbol{y}$ via the sum rule,

$$p(\Delta|\boldsymbol{y}) = \sum_c p(\Delta|c,\boldsymbol{y}) \underbrace{p(c|\boldsymbol{y})}_{w_c}. \tag{4.0.1}$$

[3] In case of an intercepts-only model, one might consider setting the prior weight on CBM to 1.0; see Prop. B.4.2.

The first term can be approximated as:

$$p(\Delta|c, \boldsymbol{y}) = \int p_c(\Delta|\boldsymbol{B}) p(\boldsymbol{B}|\boldsymbol{y}) \, d\boldsymbol{B} \approx \frac{1}{T} \sum_{t=1}^{T} p_c(\Delta|\boldsymbol{B}^t)$$

using T samples from our approximate posterior $\boldsymbol{B}^t \sim q$.

The second term $w_c = p(c|\boldsymbol{y})$ defines the posterior probability of choosing model c, which via Bayes rule is

$$w_c = \frac{p_c(\boldsymbol{y})\pi_c}{p_{\text{CBC}}(\boldsymbol{y})\pi_{\text{CBC}} + p_{\text{CBM}}(\boldsymbol{y})\pi_{\text{CBM}}}. \tag{4.0.2}$$

Each evidence term $p_c(\boldsymbol{y})$ is defined in Eq. (3.1.1a). While this term cannot be computed directly due to an intractable integral, we can estimate it via a Monte Carlo (MC) approximation of the conventional evidence lower bound defined in Eq. (3.2.1). For each model c, we estimate the evidence as:

$\log p_c(\boldsymbol{y}) \approx \frac{1}{S} \sum_{s=1}^{S} \log p_c(\boldsymbol{y}|\boldsymbol{B}^s) + D_{\text{KL}}(q(\boldsymbol{B}) \parallel \pi(\boldsymbol{B}))$, where the first term is model c's likelihood averaged over S MC samples from our approximate posterior $\boldsymbol{B}^s \sim q$, and the second KL term has a closed-form solution because our prior $\pi(\boldsymbol{B})$ and chosen variational factor $q(\boldsymbol{B})$ are exponential family distributions (Nielsen & Nock, 2010). Alternatively, we could use recent importance-sampling bounds (Burda et al., 2015) to get even more accurate approximations at the cost of increased computation.

Runtime cost of BMA. The posterior weights $w_c \in [0, 1]$ for each model type c can be computed once for each training set $\boldsymbol{y}$, at a linear cost in the number of examples N and categories K, and then stored in memory for all future uses. Then, using pre-computed weights, the cost of computing the probability of each new example's category y_* given $\boldsymbol{x}_*$ has a linear cost in K and the number of samples S.

5 Experiments

We now assess the speed and quality of our proposed CB models with IB posterior approximations. Reproducible details for all experiments are in Sec. G.

5.1 Demonstrating value of BMA for predictions

Sec. 4 described a Bayesian model averaging (BMA) technique to automatically determine the best CB target of an IB approximation for a given dataset. Here we review the effectiveness of that technique. We generated simulated datasets using $K = 3$ categories and $M = 6$ covariates at varying sizes and levels of y-from-x "predictability" (details in Sec. G.1). After fitting one approximate posterior q via IB-CAVI, we used this one posterior to make probabilistic predictions about the category labels of the held-out test set given corresponding covariates using three likelihoods: CBC only, CBM only, and our BMA approach that averages the two. Fig. 1 plots the quality of predictions as the data-to-parameter ratio changes. The first major takeaway is that our BMA averaging technique always matches the best possible prediction quality. The second takeaway is that averaging is needed: our datasets with weakly predictable outcomes favor CBM likelihoods, while those with strongly predictable outcomes favor CBC.

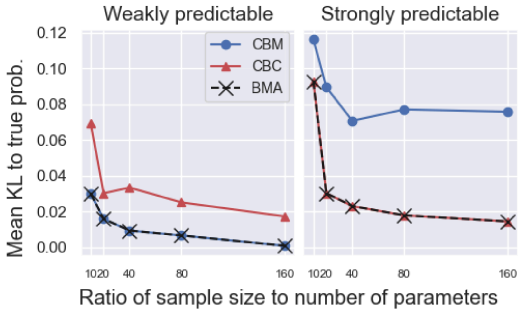

Figure 1: **Demonstrating the value of BMA prediction.** Each point corresponds to a simulated dataset ($K = 3$, $M = 6$) with number of examples N increasing along x-axis. For each dataset, we perform IB-CAVI posterior estimation then make predictions using two pure CB likelihoods, CBM-Logit and CBC-Logit, as well as Bayesian model averaging (BMA, Sec. 4) of these two models. The y-axis reports the mean KL divergence from predictions to the true probabilities, averaged across the test set (lower is better). *Left:* Categories are *weakly* predictable from covariates ($\sigma_{\text{high}} = 0.1$ in the generative process of Sec. G.1). *Right: Strong* predictability ($\sigma_{\text{high}} = 2.0$).

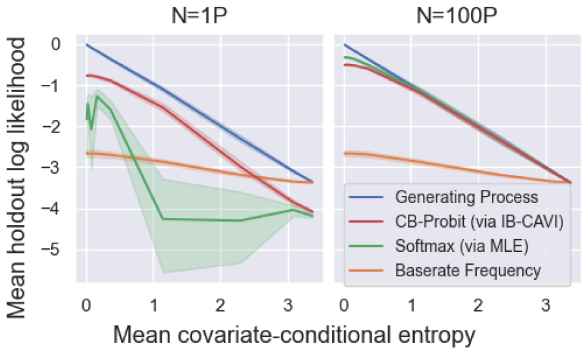

Figure 2: **Performance of IB-CAVI vs. softmax MLE given few vs. many observations-per-parameter.** Plotted is the mean holdout log likelihood as a function of mean covariate-conditional category entropy (G.1.1), which quantifies predictability. Data was simulated from a softmax regression (K=30, M=60), comparing regimes where the number of observations is modest ($N = P$, *left*) vs. abundant ($N = 100P$, *right*) relative to the number of parameters ($P = K(M + 1)$). The CB-Probit was estimated via IB-CAVI, with Bayesian model averaging of the CBC-Probit and CBM-Probit. Error bars are 95% confidence intervals for the expectation (over $D = 10$ replicated datasets) of the mean test-set log likelihood.

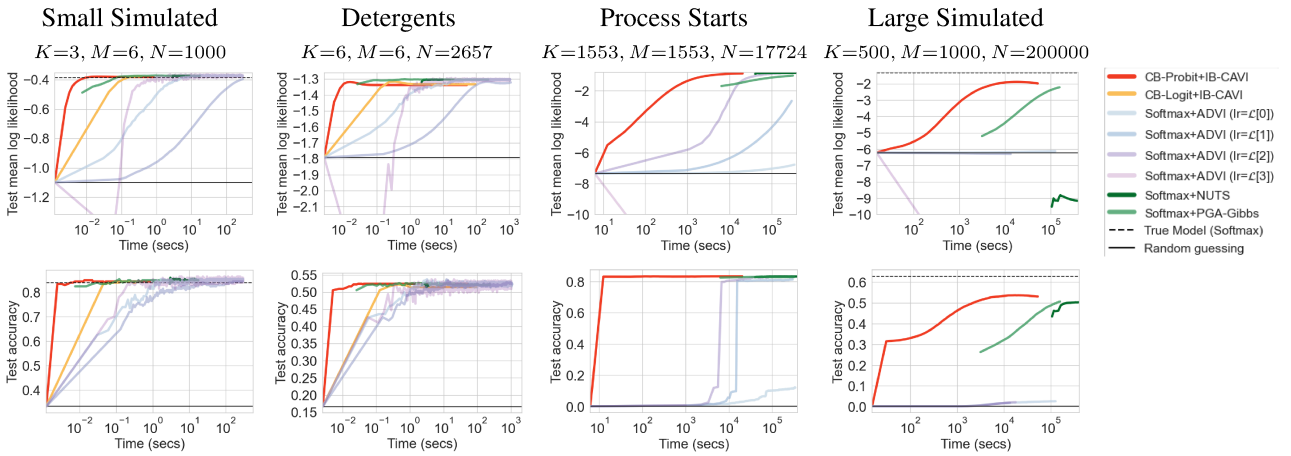

Figure 3: **Prediction quality by training time.** Bayesian inference methods are compared on real and simulated datasets with K categories, M covariates, and N instances. Prediction quality is measured by holdout log likelihood (top) and accuracy (bottom). For ADVI, we try the learning rates $\mathcal{L} := (0.01, 0.1, 1.0, 10, 100)$ recommended by (Kucukelbir et al., 2017), adjusted to $10^{-1}\mathcal{L}$ in the larger simulated dataset to reduce divergence. If a line is absent for ADVI, the method diverged. All methods were initialized at the zero matrix (corresponding to random guessing), but we do not treat the initialization as a sample for MCMC methods. Note that IB-CAVI's parallelism over K was *not* exploited here; using that, IB-CAVI's training time could be further reduced.

5.2 IB-CAVI versus maximum likelihood

Using simulated data generated by a softmax model (Sec. G.1), we assessed the quality of a CB-Probit model with IB-CAVI posterior estimation against a well-specified baseline: maximum likelihood estimation (MLE) for the softmax model. Figure 2 shows that our approach provides clear benefits over the MLE in terms of test log likelihood when the data are not sufficiently informative (i.e. the number of samples N is small relative to the number of parameters P). This result suggests that when there are few samples relative to parameters, the benefits of modeling uncertainty outweigh the cost of our approximation. When data are abundant ($N \gg P$), we expect the well-specified softmax MLE to do well, and comfortingly our prediction quality is quite close to it. This performance is assuring especially because the data was not generated by a CB model.

5.3 Prediction quality by training time

In Fig. 3, we study how different Bayesian methods for fitting categorical regressions perform on heldout test data as a function of training time. We study two real datasets, detergent choices (Imai & Van Dyk, 2005) and the computer process starts (Sec. 5.5) of one user, as well as a large and small simulated dataset generated by a softmax model (Sec. G.1). In addition to our IB-CAVI, we compared to two "gold standard" MCMC samplers for softmax models: the No U-Turn Sampler (NUTS) (Hoffman et al., 2014) and a Gibbs sampler available via Pòlya-Gamma augmentation (Polson et al., 2013). We also compared to automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017) for the softmax model.

Overall, we find that our IB-CAVI can deliver *indistinguishable accuracy* and *little-to-no cost in log likelihood* compared to alternative methods for categorical data, while requiring *far less time* to get there. Moreover, IB-CAVI is *reliable*; each update is exact and optimal, and unlike alternatives does not require correctly choosing a learning rate (as with ADVI) or tuning period length (as with NUTS).

These conclusions are corroborated on two other real datasets, glass identification (Dua & Graff, 2017) and Anuran frog calls (Colonna et al., 2017), in Sec. G.4. The heldout log likelihood plots sometimes suggest that IB-CAVI eventually overfits slightly (see *Detergents* in Fig. 3), though we can see the same behavior from Gibbs samplers (see Fig. G.2).

5.4 Assessing quality of the IB approximation

To more directly assess the quality of the approximation required by our IB-CAVI approach, we compared the predicted category probabilities to those of the true model used to generate simulated data. Table G.1 quantifies that the KL divergence from predicted to truth is always less than 0.10 across a range of dataset sizes. Histograms comparing the posterior distribution over category probabilities inferred by IB-CAVI, NUTS, and ADVI are visualized in Fig. G.3. These empirical checks, combined with our previous experiments in Sec. 5.2 and 5.3, reveal that IB approximation can be used to obtain high-quality probabilistic predictions that are competitive with more expensive truly-categorical

approaches, especially when used with Bayesian model averaging (Sec. 5.1). Future research could provide an analytical characterization of the approximation error.

5.5 Intrusion detection in a cybersecurity application

Now we show how IB-CAVI can address a real problem in which there are many categorical outcomes. User behavior analytics is a branch of cybersecurity which attempts to learn the "normal" usage patterns of users on a computer network. Deviations from normal behavior can point to suspicious or malicious activity that may be caused by unauthorized access to the network (e.g., from compromised user credentials), which we refer to as *intrusion*.

We analyze $U = 32$ users from Los Alamos National Laboratory's corporate computer network (Kent, 2015). For each user, we train a categorical GLM to learn a probability distribution over that user's process starts given previously started processes. There were $K = 1,553$ unique processes started. Our autoregressive-like featurization strategy (see Sec. G.6.2) produced $P = K(K+1) = 2.4$ million model parameters, which far exceeds the number of process starts per user ($N \approx 18,000$). This is a *small data-per-parameter* ($N << P$) regime, where Bayesian methods can show benefits over maximum likelihood point estimation, as we observed in Fig. 2 (left).

We perform *simulated intrusions*, using each user model to score holdout process start sequences from the self vs. $(U-1) = 31$ other users. The quality of each model can be summarized with an *intrusion detection score*, defined for each user $u = 1, \ldots, 32$ as

$$\texttt{ID-score}(u) = \ell_{uu} - \frac{1}{U-1}\sum_{v \neq u} \ell_{uv}$$

where ℓ_{uv} is the mean log likelihood of holdout process starts from user v when scored with the model of user u. A higher score means a better ability to distinguish the target user u from other, potentially unauthorized users.

Using these intrusion detection scores, we can compare two different variational approaches to learning categorical GLMs: IB-CAVI (applied to CB-Probit models) vs. ADVI (applied to Softmax models). We selected CB-Probit over CB-Logit due to its lower runtime costs (see Sec. 3.3). Based on the process start results in Fig. 3, we set the ADVI learning rate to 1.0, and we grant ADVI much longer training time (200 minutes per user) than IB-CAVI (20 minutes per user). Results are shown in Figure 4. We find that despite ADVI's 10-fold advantage in training time, IB-CAVI produces markedly better intrusion detection performance, matching or beating ADVI across all 32 users.

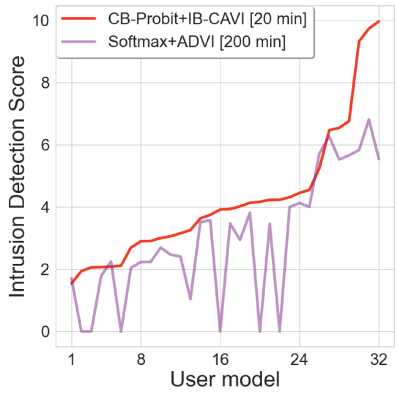

Figure 4: **Scalable intrusion detection with IB-CAVI.** We train 32 categorical models to learn the computer usage patterns of 32 users from a corporate network. Many ($K = 1,553$) unique computer processes were started. Our autoregressive-like featurization strategy produced on the order of 2.4 million ($= K^2$) parameters. Due to the large scale, we compare only variational methods. We grant IB-CAVI much shorter training time (20 minutes per user) than ADVI (200 minutes per user), yet find that IB-CAVI produces markedly better intrusion detection performance.

6 Conclusion

We have defined a class of categorical models amenable to a binary approximation that enables fast and closed-form posterior estimation. We give a likelihood bound that justifies this approximation, and demonstrate how model averaging can improve prediction quality with no additional training cost. We find that on real data the posterior estimated via our binary approximation can deliver categorical predictions of similar quality as more expensive baseline methods in a fraction of the time. Future work could provide a theoretical characterization of approximation error or determine if other constructions of CB models are useful. Our IB approximation could also be extended to more sophisticated hierarchical models, including models with variable selection priors.

Acknowledgements

This research was sponsored by the U.S. Army DEVCOM Soldier Center, and was accomplished under Cooperative Agreement Number W911QY-19-2-0003. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army DEVCOM Soldier Center, or the U.S. Government. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

We also acknowledge support from the U.S. National Science Foundation under award HDR-1934553 for the Tufts T-TRIPODS Institute. SA is supported by NSF CCF 1553075, NSF DRL 1931978, NSF EEC 1937057, and AFOSR FA9550-18-1-0465. ELM is supported by NSF grants 1934553, 1935555, 1931978, and 1937057. MCH is supported by NSF IIS-1908617.

References

Albert, J. H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Armagan, A. and Zaretzki, R. L. A note on mean-field variational approximations in bayesian probit models. *Computational statistics & data analysis*, 55(1):641–643, 2011.

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

Braun, M. and McAuliffe, J. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

Brown, P. J. and Griffin, J. E. Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis*, 5(1):171–188, 2010.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Burgette, L. F., Puelz, D., and Hahn, P. R. A symmetric prior for multinomial probit models. *Bayesian Analysis*, 16(3):991–1008, 2021.

Carvalho, C. M., Polson, N. G., and Scott, J. G. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pp. 73–80. PMLR, 2009.

Chintagunta, P. K. and Prasad, A. R. An empirical investigation of the "dynamic mcfadden" model of purchase timing and brand choice: Implications for market structure. *Journal of Business & Economic Statistics*, 16(1): 2–12, 1998.

Choi, H. M., Hobert, J. P., et al. The polya-gamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013.

Cole, D. *Parameter redundancy and identifiability*. CRC Press, 2020.

Colonna, J., Nakamura, E., Cristo, M., and Gordo, M. Anuran Calls (MFCCs). UCI Machine Learning Repository, 2017. URL https://archive-beta.ics.uci.edu/ml/datasets/anuran+calls+mfccs.

Colonna, J. G., Gama, J., and Nakamura, E. F. How to correctly evaluate an automatic bioacoustics classification method. In *Conference of the Spanish association for artificial intelligence*. Springer, 2016.

Consonni, G. and Marin, J.-M. Mean-field variational approximate bayesian inference for latent variable models. *Computational Statistics & Data Analysis*, 52(2):790–798, 2007.

Deng, J., Berg, A. C., Li, K., and Fei-Fei, L. What Does Classifying More Than 10,000 Image Categories Tell Us? In *European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Depraetere, N. and Vandebroek, M. A comparison of variational approximations for fast inference in mixed logit models. *Computational Statistics*, 32(1):93–125, 2017.

Dua, D. and Graff, C. Uci machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Durante, D. and Rigon, T. Conditionally conjugate mean-field variational bayes for logistic models. *Statistical science*, 34(3):472–485, 2019.

Evett, I. W. and Spiehler, E. J. Rule induction in forensic science. In *KBS in Government*, pp. 107–118. Online Publications, 1987. URL http://www.cs.ucl.ac.uk/staff/W.Langdon/ftp/papers/evett_1987_rifs.pdf.

Fasano, A., Durante, D., and Zanella, G. Scalable and accurate variational bayes for high-dimensional binary regression models. *arXiv preprint arXiv:1911.06743*, 2019.

Foster, I. Sec. 1.4 Parallel Algorithm Examples. In *Designing and Building Parallel Programs*. Addison-Wesley, 1995. URL http://www.mcs.anl.gov/~itf/dbpp/text/book.html.

Galy-Fajou, T., Wenzel, F., Donner, C., and Opper, M. Multi-class gaussian process classification made conjugate: Efficient inference via data augmentation. In *Uncertainty in Artificial Intelligence*, pp. 755–765. PMLR, 2020.

German, B. Glass identification. UCI Machine Learning Repository, 1987. URL https://archive.ics.uci.edu/ml/datasets/glass+identification.

Held, L. and Holmes, C. C. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, 2006.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Hoffman, M. D., Gelman, A., et al. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Holsclaw, T., Greene, A. M., Robertson, A. W., Smyth, P., et al. Bayesian nonhomogeneous markov models via pólya-gamma data augmentation with applications to rainfall modeling. *The Annals of Applied Statistics*, 11 (1):393–426, 2017.

Hughes, M. C. and Sudderth, E. Memoized online variational inference for dirichlet process mixture models. *Advances in Neural Information Processing Systems*, 26: 1133–1141, 2013.

Imai, K. and Van Dyk, D. A. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124(2):311–334, 2005. URL https://github.com/kosukeimai/MNP/blob/master/data/detergent.txt.gz.

Jaakkola, T. S. and Jordan, M. I. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

Johndrow, J., Dunson, D., and Lum, K. Diagonal orthant multinomial probit models. In *Artificial Intelligence and Statistics*, pp. 29–38. PMLR, 2013.

Kent, A. D. Comprehensive, multi-source cyber-security events data set. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2015. URL https://csr.lanl.gov/data/cyber1/.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of machine learning research*, 2017.

Linderman, S. W., Johnson, M. J., and Adams, R. P. Dependent multinomial models made easy: Stick breaking with the pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, 2015:3456–3464, 2015.

Madigan, D., Raftery, A. E., Volinsky, C., and Hoeting, J. Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR*, pp. 77–83, 1996.

Makalic, E. and Schmidt, D. F. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23 (1):179–182, 2015.

Murphy, K. P. Sec. 14.5.2.4 SVMs for multi-class classification. In *Machine Learning: A Probabilistic Perspective*, pp. 534. MIT press, 2012.

Nielsen, F. and Nock, R. Entropies and cross-entropies of exponential families. In *2010 IEEE International Conference on Image Processing*, pp. 3621–3624. IEEE, 2010.

Ormerod, J. T. and Wand, M. P. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011. URL http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Polson, N. G., Scott, J. G., and Windle, J. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

Rifkin, R. and Klautau, A. In Defense of One-Vs-All Classification. *Journal of Machine Learning Research*, 5, 2004.

Turcotte, M., Moore, J., Heard, N., and McPhall, A. Poisson factorization for peer-based anomaly detection. In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pp. 208–210. IEEE, 2016.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, C. and Blei, D. M. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(4), 2013.

Wojnowicz, M., Aeron, S., Miller, E., and Hughes, M. C. Easy variational inference for categorical observations via a new view of diagonal orthant probit models. In *The 4th Workshop on Tractable Probabilistic Modeling*, 2021.

Zhang, Q. and Zhou, M. Permuted and augmented stick-breaking bayesian multinomial regression. *The Journal of Machine Learning Research*, 18(1):7479–7511, 2017.

Appendix Contents

Code and Data Availability

Open-source python code for reproducing experiments can be found at `https://github.com/tufts-ml/categorical-from-binary`

The above repository also contains code to download the simulated datasets and open-access real datasets (Detxergent, Glass, and Frog Calls, Cybersecurity events) used in this work, along with any preprocessing used.

A Background: General binary regression

Here we obtain an additional interpretation for the "success" probability in a general binary regression model (Eq. (2.2.1)) if we are willing to make a few additional assumptions. When H is the cdf of a distribution $\mathcal{H}$ that is a location family with a symmetric density (so $\mathcal{H}$ could be Gaussian, Logistic, Cauchy, Laplace, Student-t, and so on), a simple argument reveals the further interpretation that $\mathbb{P}(\breve{y}_i = 1 \mid \breve{\boldsymbol{\beta}}) = \mathbb{P}\{\text{drawing a positive value from } \mathcal{H} \text{ when its mean is } \boldsymbol{x}_i^T \breve{\boldsymbol{\beta}}\}$. We formalize this in Prop. A.0.1.

Proposition A.0.1. *Let $\mathcal{H}_\mu$ be a symmetric location-scale family of probability measures that has a density h_μ which is continuous almost everywhere with respect to Lebesgue measure. Let $\mathcal{H}_\mu$ have expected value μ and variance fixed to 1. Let $z \sim \mathcal{H}_\mu$, and denote the cdf of $\mathcal{H}_\mu$ by H_μ. Then $H_0(\mu) = \mathbb{P}(z \geq 0)$.*

Proof.

$$\begin{aligned}
\mathbb{P}(z \geq 0) &= 1 - H_\mu(0) \\
&= \int_0^\infty h_\mu(x)\, d\,x && \text{by Riemann integrability} \\
&= \int_0^\infty h_0(x-\mu)\, d\,x && \text{as a location-scale family with unit variance} \\
&= \int_{-\mu}^\infty h_0(u)\, d\,u && \text{by substituting } u = x - \mu \\
&= 1 - H_0(-\mu) \\
&= H_0(\mu). && \text{by symmetry}
\end{aligned}$$

□

B Categorical-from-binary (CB) models

B.1 Impossibility of exactness in the likelihood bound

Proposition B.1.1. *Equality cannot be achieved in the likelihood bound of Eq.* (2.3.1).

Proof. By way of contradiction, suppose that

$$p_{\text{CB}}(y_i = k \mid \boldsymbol{B}) = p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_k \mid \widehat{\boldsymbol{B}} = \boldsymbol{B}) \tag{B.1.1}$$

Summing these terms over $k = 1, \ldots, K$ must yield a value of 1, since the left hand side is a probability distribution over the K categories. This says that $p_{\text{IB}}(\Upsilon \mid \boldsymbol{B}) = 1$, where we define $\Upsilon \subsetneq \{0,1\}^K$ as the space of one-hot encoded vectors. So by countable additivity,

$$p_{\text{IB}}(\Upsilon^c \mid \boldsymbol{B}) = 0 \tag{B.1.2}$$

At the same time, we must have that

$$0 < p_{\text{IB}}(\boldsymbol{v} \mid \boldsymbol{B}) < 1, \quad \forall \boldsymbol{v} \in \{0,1\}^K \tag{B.1.3}$$

Eq. (B.1.3) follows immediately from Eq. (2.2.3), because a cdf H satisfies that $H(x), 1 - H(x) \in (0,1)$ for any finite real-valued x, and $p_{\text{IB}}(\boldsymbol{v} \mid \boldsymbol{B})$ is the product of K such terms.

Now Eqs. (B.1.3) and (B.1.2) contradict, so the hypothesis in Eq. (B.1.1) is false. □

B.2 The CBC model as a normalized odds model

In the main paper, the CBC model was given as:

$$p_{\text{CBC}}(y_i{=}k \mid \boldsymbol{B}) = \frac{H(\eta_{ik}) \prod_{j \neq k} (1 - H(\eta_{ij}))}{\sum_{\ell=1}^K H(\eta_{i\ell}) \prod_{j \neq \ell} (1 - H(\eta_{ij}))} \tag{B.2.1}$$

where $\eta_{ik} = \boldsymbol{x}_i^T \boldsymbol{\beta}_k$.

A CBC model has an alternate expression as a *normalized odds* model:

$$p_{\text{CBC}}(y_i = k \mid \boldsymbol{B}) = \frac{H(\eta_{ik}) / \big(1 - H(\eta_{ik})\big)}{\sum_{\ell=1}^K H(\eta_{i\ell}) / \big(1 - H(\eta_{i\ell})\big)} \tag{B.2.2}$$

Here we show that the two representations for the CBC model (Eqs. (B.2.1) and (B.2.2)) are equal. Observe

$$\begin{aligned}
p(y_i = k \mid \boldsymbol{B}) &\overset{(2.4.2)}{=} \frac{H(\eta_{ik}) \prod_{j \neq k} \big(1 - H(\eta_{ij})\big)}{\sum_{\ell=1}^K H(\eta_{i\ell}) \prod_{j \neq \ell} \big(1 - H(\eta_{ij})\big)} \\
&\overset{1}{=} \frac{H(\eta_{ik}) \prod_{j \neq k} \big(1 - H(\eta_{ij})\big)}{\sum_{\ell=1}^K \frac{H(\eta_{i\ell})}{\big(1 - H(\eta_{i\ell})\big)} \prod_{j=1}^K \big(1 - H(\eta_{ij})\big)} \\
&\overset{2}{=} \frac{H(\eta_{ik}) / \big(1 - H(\eta_{ik})\big)}{\sum_{\ell=1}^K H(\eta_{i\ell}) / \big(1 - H(\eta_{i\ell})\big)}
\end{aligned}$$

Equality (1) just multiplies by $1 = \frac{a}{a}$ in the numerator and denominator, and Equality (2) just cancels.

B.3 Membership of CBC and CBM models in the CB class.

Proposition B.3.1. CBC *and* CBM *models are categorical-from-binary (*CB*) models.*

Proof. We begin with CBM models. For any $k = 1, ..., K$,

$$p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_k \mid \widehat{\boldsymbol{B}} = \boldsymbol{B}) < p_{\text{CBM}}(y_i = k \mid \boldsymbol{B})$$

holds if

$$w_k \prod_{j \neq k} (1 - w_j) < \frac{w_k}{\sum_{\ell=1}^K w_\ell} \tag{B.3.1}$$

for $w_k \in (0,1), k = 1, ..., K$. The implication follows from the IB and CB likelihood equations (Eqs. (2.4.1) and (2.2.3)) and the fact that any cdf H maps into $(0,1)$. Some algebra reduces Eq. (B.3.1) to

$$\left(\sum_{\ell=1}^K w_\ell\right) \left(\prod_{j \neq k} (1 - w_j)\right) < 1 \tag{B.3.2}$$

We establish Eq. (B.3.2) by induction on K. Without loss of generality, we assume $k = 1$.

- *(Base.)* We show Eq. (B.3.2) holds for $K = 2$:

$$w_1(1 - w_2) + w_2(1 - w_2) \overset{(w_1<1)}{<} (1 - w_2)(1 + w_2) \overset{(w_2<1)}{<} 1.$$

- *(Step.)* We show that if Eq. (B.3.2) holds for some K then it holds for $K + 1$. We define $\alpha_{k,K} := \sum_{\ell=1}^K w_\ell \prod_{j \neq k} (1 - w_j)$ and assume $\alpha_{k,K} < 1$. Now

$$\begin{aligned}
\alpha_{k,K+1} &= \underbrace{\alpha_{k,K}}_{<1 \text{ by hypoth.}} (1 - w_{K+1}) + w_{K+1} \underbrace{\prod_{j \neq k} (1 - w_j)}_{<1} \\
&< (1 - w_{K+1}) + w_{K+1} = 1.
\end{aligned}$$

Now we proceed to CBC models. We must show that for any $k = 1, ..., K$,

$$p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_k \mid \widehat{\boldsymbol{B}} = \boldsymbol{B}) < p_{\text{CBC}}(y_i = k \mid \boldsymbol{B}) \quad \text{(B.3.3)}$$

By the likelihood equations for the IB and CBC models (Eqs. (2.2.3) and (2.4.2)), we can write

$$p_{\text{CBC}}(y_i = k \mid \boldsymbol{B}) = \frac{p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_k \mid \widehat{\boldsymbol{B}} = \boldsymbol{B})}{p_{\text{IB}}(\Upsilon \mid \widehat{\boldsymbol{B}} = \boldsymbol{B})}, \quad \text{(B.3.4)}$$

where $\Upsilon \subset \{0,1\}^K$ is the space of one-hot encoded vectors. Now take any $\boldsymbol{v} \in \Upsilon^c$. Then

$$p_{\text{IB}}(\Upsilon \mid \boldsymbol{B}) \overset{subadditivity}{\leq} p_{\text{IB}}(\{0,1\}^K \mid \boldsymbol{B}) - p(\boldsymbol{v} \mid \boldsymbol{B}) \overset{Eq.\ (B.1.3)}{<} p_{\text{IB}}(\{0,1\}^K \mid \boldsymbol{B}) = 1.$$

So the denominator of Eq. (B.3.4) is non-negative and less than one, which implies Eq. (B.3.3). □

B.4 Impossibility of exact inference on a CBM or CBC model via inference on an IB model

As discussed in Sec. 2.1, efficient Bayesian inference exists for IB models. As a result, we would like to relate CB models to IB models in such as way as to obtain efficient inference for the *categorical* models.

Here we consider whether an *exact* relation exists. In Sec. B.4.1, we show that whereas CBC and CBM models are non-identifiable, IB models are identifiable (see Def. B.4.1). Now suppose there were an identifiability constraint such that a CB model under the identifiability constraint provided the same probabilities as an IB model applied to one-hot-encoded category representations. Then there would be no need to consider *approximate* inference, as inference on the regression weights $\widehat{\boldsymbol{B}}$ for the IB model would be *exactly* equivalent to inference on the regression weights for the (identified) CB model.

Unfortunately, no such identifiability constraint exists for either CBC or CBM models. While previous work (Johndrow et al., 2013) has suggested that the CBC model can be identified via the IB model, Prop. B.4.2 shows that this cannot be true. On the other hand, while Prop. B.4.2 may elicit hope that the CBM model could be identified via an IB model, Sec. B.5.1 reveals empirically that this is no longer true once covariates are introduced.

B.4.1 Can CB models be identified via IB models?

Let us begin with a general definition of identifiability.

Definition B.4.1. [*(Cole, 2020) pp.35*] A likelihood $p(\cdot \mid \theta)$ with parameter θ is globally identifiable if $p(\cdot \mid \theta_1) = p(\cdot \mid \theta_2)$ implies that $\theta_1 = \theta_2$. A model is locally identifiable if there exists an open neighborhood in the parameter space of θ such that this is true. Otherwise a model is non-identifiable. △

For models considered in the main body of this paper, identifiability requires

$$p_m(\cdot \mid \boldsymbol{B}_1) = p_m(\cdot \mid \boldsymbol{B}_2) \implies \boldsymbol{B}_1 = \boldsymbol{B}_2 \quad \text{(B.4.1)}$$

where m can take the value of either IB, CBC, or CBM models.

We now investigate the identifiability of these models in the simplest possible scenario: the intercepts-only (no covariates) setting. In this setting, $M = 1$, $\boldsymbol{x}_i = 1$ for all $i = 1, 2, \ldots, N$, and the regression matrix simplifies to a vector $\boldsymbol{\beta} \in \mathbb{R}^K$. Although this setting is simple, it is sufficient to provide some insight about identifiability.

Proposition B.4.1. *The* CBC *and* CBM *models are non-identifiable in the intercepts-only setting.*

Proof. Let $(p_1, \ldots, p_K)$ be a probability mass function over K categories. Then we can construct regression weights $\boldsymbol{\beta}^{\text{CBM}}, \boldsymbol{\beta}^{\text{CBC}} \in \mathbb{R}^K$ for the two models by taking the kth entry of each vector to be given by

$$\beta_k^{\text{CBM}} \in \left\{ H^{-1}(rp_k) \right\}_{r \in (0, \min_\ell 1/p_\ell)} \quad \text{(B.4.2)}$$

$$\beta_k^{\text{CBC}} \in \left\{ H^{-1}\left(\frac{sp_k}{1 + sp_k} \right) \right\}_{s > 0} \quad \text{(B.4.3)}$$

where Eq. (B.4.3) follows from setting the model's category probabilities in (B.2.2) to p_k. Therefore Eq. (B.4.1) is not satisfied for any open neighborhood of the parameter space. □

Now let us consider the IB model, specifically in the case where we use it to do inference with one-hot encoded representations of categorical data, $\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_{y_i}$ for $y_i \in \{1, \ldots, K\}$. If we set

$$\begin{aligned} p_{\text{IB}}(1, 0, 0 \ldots, 0) &= p_1 \\ p_{\text{IB}}(0, 1, 0 \ldots, 0) &= p_2 \\ &\vdots \\ p_{\text{IB}}(0, 0, 0 \ldots, 1) &= p_K \end{aligned}$$

and $p_{\text{IB}}(\boldsymbol{v}) = 0$ for all vectors $\boldsymbol{v} \in \{0,1\}^K$ that are not one-hot, then we have transformed $(p_1, \ldots, p_K)$ into a probability mass function over K-bits. Via Eq. (2.2.2), this pmf implies a unique vector of regression weights $\widehat{\boldsymbol{\beta}}^{\text{IB}} \in \mathbb{R}^K$, with kth entry given by

$$\widehat{\beta}_k^{\text{IB}} = H^{-1}(p_k) \quad \text{(B.4.4)}$$

So unlike the CBC and CBM models, the IB model *is* identifiable (and globally so), at least when we apply it only to one-hot encoded data.

Remark B.4.1. If we consider the special case where $p_1, ..., p_K$ are the empirical category frequencies in a K-class intercepts-only dataset, $p_k = \frac{1}{N} \sum_{i=1}^N 1_{y_i = k}$, then Eqs. (B.4.2), (B.4.3), and (B.4.4) give the maximum likelihood estimators. Note in particular that since CBC and

CBM models are non-identifiable, the maximum likelihood estimators are not unique. △

So whereas the CBC and CBM models are non-identifiable, the IB model is globally identifiable. Moreover, CBC and CBM models are "built from" IB models in the sense described in Sec. 2.4. These observations give rise to a natural question. Does the IB model give an *identifiability constraint* for the CBC or CBM models? That is, given a set of regression weights $\boldsymbol{B}$ that produce an equivalent likelihood for the categorical model, can we choose a representative by setting $\boldsymbol{B} = \widehat{\boldsymbol{B}}_{\text{MLE}}$, where $\widehat{\boldsymbol{B}}_{\text{MLE}}$ is the maximum likelihood estimate of the regression weights of an IB model given one-hot encoded representations of categorical data? We address this question mathematically in Prop. B.4.2 for the intercepts-only case, and empirically in Sec. B.5.1 for the with-covariates case.

Proposition B.4.2. *In the intercepts-only setting, the* IB *model gives an identifiability constraint for the* CBM *model, but not for the* CBC *model.*

Proof. For the CBM model, take r=1 in Eq. (B.4.2), and we obtain (B.4.4). For the CBC model, we require

$$H^{-1}(p_k) \in \left\{ H^{-1}\left(\frac{sp_k}{1+sp_k}\right) \right\}_{s>0}$$
$$\implies \exists s > 0 : p_k = \frac{s-1}{s} \implies p_k = 1/K.$$

Since the empirical probabilities are not always uniform, the IB model cannot provide an identifiability constraint for the CBC model. □

Remark B.4.2. From Prop. B.4.2, it follows that in the intercepts-only setting, $\widehat{\boldsymbol{B}}_{\text{MLE}}$, the maximum likelihood estimate of the regression weights of an IB model given one-hot encoded representations of categorical data, is an MLE for the CBM model, but not for the CBC model. △

B.5 Evaluating CB models as targets of an IB approximation

In this section, we evaluate two different classes of CB models (CBC and CBM) in terms of how well they serve as targets of an IB approximation. In Sec B.5.1, we make an evaluation in terms of "soft" predictions (i.e. likelihood). In Sec. B.5.2, we make an evaluation in terms of "hard" predictions (i.e. misclassification rates).

B.5.1 Soft predictions

Now we evaluate the quality of CBC and CBM as targets of an IB approximation with respect to *soft predictions* (the probability vectors produced by a categorical likelihood).

Methodology. We generate two datasets using the data generation technique of Section G.1.1, fixing in both the response predictability at $\sigma_{\text{high}} = 0.1$ to create a challenging problem. The smaller dataset has $K{=}3, M{=}1, N{=}1800$. The larger dataset has $K{=}20, M{=}40, N{=}12000$.

For each dataset, we perform gradient descent (using automatic differentiation of the relevant likelihoods from multiple initialization seeds) to estimate $\widehat{\boldsymbol{B}}^{\text{CBC}}_{\text{MLE}}$ and $\widehat{\boldsymbol{B}}^{\text{CBM}}_{\text{MLE}}$, the MLEs for the CBC-Probit and CBM-Probit models, respectively. Similarly, we estimate $\widehat{\boldsymbol{B}}^{\text{IB}}_{\text{MLE}}$, the MLE for the IB-Probit model when it is fit on one-hot encoded representations of the categorical data.

Then, for each example i in the training data, we compute the categorical probability vector $\boldsymbol{s}_i \in \Delta_{K-1}$ using the weights that minimize the truly-categorical CB likelihood, as well as probability vector $\hat{\boldsymbol{s}}_i \in \Delta_{K-1}$ corresponding to the weights estimated to minimize the IB likelihood. (Recall that vector $\boldsymbol{s}_i$ can be produced given weights $\boldsymbol{B}$ via Eq. (1.1.1)). Suppose k is the class with the highest probability in the vector $\hat{\boldsymbol{s}}_i$: we wish to compare the signed error between the "ideal" $\boldsymbol{s}_{ik}$ and our approximations $\hat{\boldsymbol{s}}_{ik}$ in order to assess the quality of our IB approximation.

Results. Figure B.1 demonstrates two important points:

1. The MLE for the IB model is not an *exact* MLE for the either the CBC or CBM models. As a result, neither of the two models is a globally identified IB model. (If they were, then the signed error would always equal 0.) This provides an empirical refutation that IB gives an identifiability constraint for CB models. Previously, Johndrow et al. (2013) suggested that such a relationship might hold (at least for CBC models). Prop. B.4.2 proved that such a relationship cannot hold for CBC models in the intercepts-only setting ($M = 0$), and these results provide an empirical refutation for CBC models in the with-covariates setting ($M \geq 1$). Moreover, while Prop. B.4.2 revealed that the identification relationship *does* hold for CBM models in the intercepts-only setting ($M = 0$); it apparently does not hold in general.
2. Neither CBC models nor CBM models are uniformly dominant as a target of an IB approximation. For the smaller dataset, the CBM model is superior to CBC; the approximation error is lower in the top right panel than the top left panel. However, for the larger dataset, the CBC model is superior to CBM; the approximation error is lower in the bottom left panel than the bottom right panel. Thus, *either* of the $\{\text{CBC}, \text{CBM}\}$ models can provide superior performance to the other as targets of an IB approximation. The relative advantage depend on properties of the data.

Figure B.1 also provides information on the *amount* of error incurred by using an IB approximation. For similar results in the context of IB-CAVI algorithm, see Table G.1.

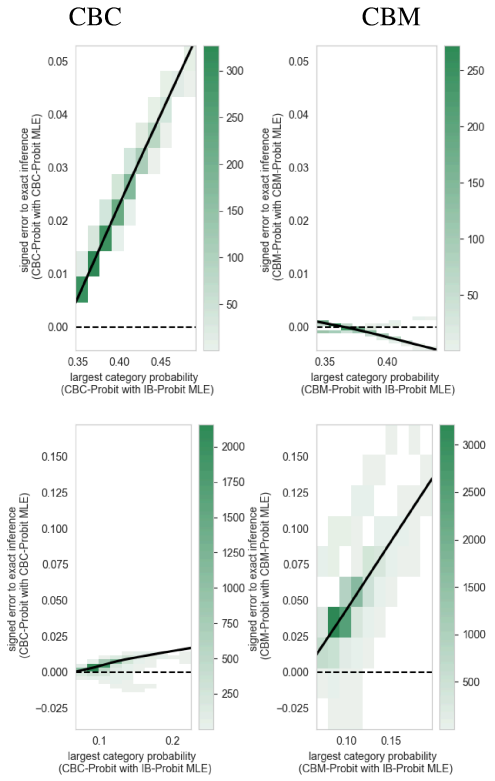

Figure B.1: 2D histograms of signed error $s_{ik} - \hat{s}_{ik}$ in the probability mass assigned to the most probable category k by a CB model and its IB approximation, using ML estimated weights. Each panel shows a distribution across training examples, where each example i contributes to the bin corresponding to the probability of its most probable category k (x-axis) and its signed error (y-axis). Darker colors imply more examples belong to that bin. *Top panels:* Smaller dataset with $K = 3, M = 1, N = 1800, \sigma^2_{high} = 0.1$. *Bottom panels:* Larger dataset with $K = 20, M = 40, N = 12,000, \sigma^2_{high} = 0.1$. *Left panels:* A CBC model is used to estimate the weights that determine $\boldsymbol{s}_i$. *Right panels:* A CBM model is used to estimate the weights that determine $\boldsymbol{s}_i$. The solid black line gives nonparametric lowess (locally weighted linear regression) estimates of the expected relationship between largest category probability and error.

B.5.2 Hard predictions

With respect to *hard predictions* (misclassification rates), the quality of the IB approximation strategy is invariant to whether CBC or CBM is used as the target of the approximation. This fact is implied by the proposition below.

Proposition B.5.1. *For fixed regression weights and covariates, the likelihoods for* CBC *and* CBM *put the most probability mass on the same category. That is*

$$\begin{aligned} k^* &:= \operatorname*{argmax}_k p_{\text{CBC}}(y_i = k \mid \boldsymbol{\beta}) \\ &= \operatorname*{argmax}_k p_{\text{CBM}}(y_i = k \mid \boldsymbol{\beta}). \end{aligned}$$

Moreover, the most likely category matches that given by the most likely one-hot vector $\boldsymbol{e}_k$ from the IB model:

$$k^* = \operatorname*{argmax}_k p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_k \mid \boldsymbol{\beta}).$$

Proof. Define the linear predictor $\eta_{ik} := \boldsymbol{x}_i^T \boldsymbol{\beta}_k$. Then

$$\begin{aligned} p_{\text{CBM}}(y_i = k \mid \boldsymbol{\beta}) &\overset{(2.4.1)}{=} \frac{H(\eta_{ik})}{\sum_{k=1}^K H(\eta_{ik})} \\ &\propto f_{\text{CBM}}(\eta_{ik}) \\ p_{\text{CBC}}(y_i = k \mid \boldsymbol{\beta}) &\overset{(B.2.2)}{=} \frac{H(\eta_{ik})/\big(1 - H(\eta_{ik})\big)}{\sum_{k=1}^K H(\eta_{ik})/\big(1 - H(\eta_{ik})\big)} \\ &\propto f_{\text{CBC}}(\eta_{ik}) \\ p_{\text{IB}}(\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_k \mid \boldsymbol{\beta}) &\overset{(2.2.3),\,(B.2.2)}{=} H(\eta_{ik})/\big(1 - H(\eta_{ik})\big) \\ &\propto f_{IB}(\eta_{ik}) \end{aligned}$$

where we have introduced notation f_{CBM}, f_{CBC}, f_{IB} to refer to the potential functions (unnormalized category probabilities) for each model. Now since H is an increasing function, f_{CBC}, f_{CBM} f_{IB} are all increasing functions of the linear predictor η_{ik}. Thus,

$$\begin{aligned} \operatorname*{argmax}_k p_{\text{CBC}}(y_i = k \mid \boldsymbol{\beta}) &= \operatorname*{argmax}_k p_{\text{CBM}}(y_i = k \mid \boldsymbol{\beta}) \\ &= \operatorname*{argmax}_k p_{\text{IB}}(\widehat{\boldsymbol{y}} = \boldsymbol{e}_k \mid \boldsymbol{\beta}) \\ &= \operatorname*{argmax}_k \eta_{ik} \end{aligned}$$

and the proposition holds. □

Remark B.5.1. Proposition B.5.1 can be extended to a more general proposition that the two approximation strategies provide identical rankings for *all* categories. Indeed, we observe this phenomenon in practice. △

C General variational algorithm for CB models

Here provide a strategy for performing variational inference with *any* model $\mathcal{M}$ whose joint density includes a CB likelihood; that is, the joint density has the form

$$p_M(\boldsymbol{y}, \boldsymbol{u}) = p_{\text{CB}}(\boldsymbol{y} \mid \boldsymbol{u})\pi(\boldsymbol{u})$$

where $\boldsymbol{y}$ are observed random variables, $\boldsymbol{u} = (u_v)_{v=1}^V$ are unobserved random variables, p_{CB} is a CB likelihood, and π is a prior density. This strategy can be applied to the CBC and CBM models, but it is also extensible to more complicated graphical models, such as hierarchical models or models with variable selection priors (such as the normal-gamma prior (Brown & Griffin, 2010) which generalizes Bayesian lasso, or the horseshoe prior (Carvalho et al., 2009), for which a conditionally conjugate formulation exists (Makalic & Schmidt, 2015)). We summarize our VI strategy in Algorithm 1.

Algorithm 1: `IB-CAVI` *for approximate Bayesian inference on any model with a* CB *likelihood.* Given N observed categorical responses $\boldsymbol{y} \in \{1, ..., K\}^N$ and covariates $\boldsymbol{X} \in \mathbb{R}^{N \times M}$

1. Form the matrix of one-hot encoded responses $\boldsymbol{E_y} = (\boldsymbol{e}_{y_1}^T, \ldots, \boldsymbol{e}_{y_N}^T)^T \in \mathbb{R}^{N \times K}$.

2. Take the IB model $\mathcal{M}_{\text{IB}}$ which is the *base* (see Sec. 2.3) of the CB model, and use it to compute *surrogate complete conditionals*: $\{\log p_{\mathcal{M}_{\text{IB}}}(u_v \mid \boldsymbol{u}_{-v}, \boldsymbol{E_y})\}$.

3. Take $\mathcal{Q}$ to be a mean-field family with factorization: $q(u_1, ..., u_V) = \prod_{v=1}^V q_v(u_v)$. (The regression weights $\widehat{\boldsymbol{B}}$ are included in this set, as are auxillary variables $\widehat{\boldsymbol{Z}}$ if augmentation is used. If the CB likelihood is embedded within a more complicated graphical model, there may be other unobserved random variables as well.)

4. Define the objective: $\mathcal{L}_{\mathcal{M}}(q) = \mathbb{E}_q[\log \frac{p_{\mathcal{M}_{\text{IB}}}(\boldsymbol{E_y}, \boldsymbol{u})}{q(\boldsymbol{u})}]$.

5. Optimize $\mathcal{L}_{\mathcal{M}}(q)$ using optimal coordinate ascent updates (Blei et al., 2017)): $q_v(u_v) \propto \exp\left\{\mathbb{E}_{q_{-v}}\left[\log p_{\mathcal{M}_{\text{IB}}}(u_v \,|\, \boldsymbol{u}_{-v}, \boldsymbol{E_y})\right]\right\}$. If the complete conditional is an exponential family with natural parameter η_v, so is its optimal update, with natural parameter given by

$$\nu_v = \mathbb{E}_{q_{-v}}[\eta_v(\boldsymbol{u}_{-v}, \boldsymbol{E_y})] \tag{C.0.1}$$

The updates in Algorithm 1 will yield a density q^* that is a local maximum of the `ELBO` of the surrogate model (Ormerod & Wand, 2010), and therefore a local maximum of a surrogate bound on the intended truly categorical model. For a CB model with a probit or logit link, a Gaussian prior on the regression weights, and use of appropriate augmentation, all conditionals enjoy closed-form updates in Eq. (C.0.1), and the objective function $\mathcal{L}_{\mathcal{M}}$ is also available in closed-form, which is useful for convergence monitoring. For details, see Secs. D and E.

D Variational inference for CB-Probit Models

Here we present closed-form variational inference for CB-Probit models. The inference follows naturally from our IB-CAVI procedure in Algorithm 1.

D.1 Distributional preliminaries

D.1.1 Entropy facts about Multivariate Gaussian

If p, q are the densities of two different d-variate Gaussian distributions with parameters $\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p$ and $\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q$, respectively, then the entropy is given by

$$\mathbb{H}[q] = \frac{1}{2} \log \left[(2\pi e)^d |\boldsymbol{\Sigma}_q| \right] \tag{D.1.1}$$

The KL divergence is given by

$$\begin{aligned} \text{KL}(q \,||\, p) = \frac{1}{2} \Big[\log \frac{|\boldsymbol{\Sigma}_p|}{|\boldsymbol{\Sigma}_q|} - d \\ + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^T \boldsymbol{\Sigma}_p^{-1} (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) + \text{tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_q) \Big] \end{aligned} \tag{D.1.2}$$

The cross-entropy of two multivariate Gaussians can then be determined from (D.1.1) and (D.1.2) via the relation

$$\mathbb{H}[q, p] = \mathbb{H}[q] + \text{KL}\big(q \,\big|\big|\, p\big) \tag{D.1.3}$$

D.1.2 Univariate normals truncated to positive or negative reals

The univariate truncated normal distribution $\mathcal{TN}(\mu, \sigma^2, \Upsilon)$ results when a normal distribution $\mathcal{N}(\mu, \sigma^2)$ is truncated to some set $\Upsilon \subseteq \mathbb{R}$. Note that the parameters μ, σ^2 denote the mean and variance of the *parent* normal distribution; i.e. if $X \sim \mathcal{TN}(\mu, \sigma^2, \Upsilon)$ then $\mathbb{E}[X] \neq \mu$ (unless $\Upsilon = \mathbb{R}$).

If we assume that the truncation set is an interval $\Upsilon = (a, b)$ for $a, b \in \mathbb{R}$, then the distribution $\mathcal{TN}(\mu, \sigma^2, (a, b))$ has p.d.f.

$$f(x; \mu, \sigma^2, a, b) = \frac{\phi_{\mu,\sigma^2}(x)}{\Phi_{\mu,\sigma^2}(b) - \Phi_{\mu,\sigma^2}(a)} 1_{a \leq x \leq b}(x)$$

where ϕ_{μ,σ^2} and Φ_{μ,σ^2} denote the pdf and cdf, respectively, of a univariate normal distribution with mean μ and variance σ^2.

We will work with distributions truncated to the positive or negative reals, and so we define special notation: $\mathcal{N}_+(\mu, \sigma^2) := \mathcal{TN}(\mu, \sigma^2, [0, \infty))$ and $\mathcal{N}_-(\mu, \sigma^2) := \mathcal{TN}(\mu, \sigma^2, (-\infty, 0))$. In particular, we will work with random variables of the form $T_+ \sim \mathcal{N}_+(\mu, 1)$ and $T_- \sim \mathcal{N}_-(\mu, 1)$. Based on this construction, it is straightforward

to derive

$$f_{T_+}(x) = \frac{\phi(x-\mu)}{1-\Phi(-\mu)} 1_{x\geq 0}, \quad f_{T_-}(x) = \frac{\phi(x-\mu)}{\Phi(-\mu)} 1_{x<0}$$

$$\mathbb{E}[T_+] = \mu + \frac{\phi(-\mu)}{1-\Phi(-\mu)}, \quad \mathbb{E}[T_-] = \mu - \frac{\phi(-\mu)}{\Phi(-\mu)} \tag{D.1.4}$$

$$\text{Var}[T_+] = 1 - \mu(\mathbb{E}[T_+] - \mu) - (\mathbb{E}[T_+] - \mu)^2 \tag{D.1.5}$$

$$\text{Var}[T_-] = 1 - \mu(\mathbb{E}[T_-] - \mu) - (\mathbb{E}[T_-] - \mu)^2 \tag{D.1.6}$$

$$\mathbb{H}[T_+] = \ln\left(\sqrt{2\pi e}\,[1-\Phi(-\mu)]\right) - \frac{\mu\phi(-\mu)}{2(1-\Phi(-\mu))} \tag{D.1.7}$$

$$\mathbb{H}[T_-] = \ln\left(\sqrt{2\pi e}\,\Phi(-\mu)\right) + \frac{\mu\phi(-\mu)}{2\Phi(-\mu)} \tag{D.1.8}$$

where we use ϕ and Φ to refer to the pdf and cdf, respectively, of the standard normal distribution, and where $\mathbb{H}[X] = -\int f(x)\ln f(x)\,dx$ represents the differential entropy of a random variable X with density f.

Remark D.1.1. (*Representation in terms of perturbation of parent mean*) It is sometimes convenient to express the expectation of a truncated random variable as a perturbation of the expectation of its parent (pre-truncated) Gaussian random variable. To this end, for $T_s \in \{T_+, T_-\}$, we write

$$\mathbb{E}[T_s] = \mu + \delta_s(\mu), \quad \delta_s(\mu) := \begin{cases} \dfrac{\phi(-\mu)}{1-\Phi(-\mu)}, & s = + \\ -\dfrac{\phi(-\mu)}{\Phi(-\mu)}, & s = - \end{cases} \tag{D.1.9}$$

which holds by Eq. (D.1.4). △

Remark D.1.2. (*Second moments*) For $T_s \in \{T_+, T_-\}$, we have

$$\begin{aligned} \mathbb{E}[T_s^2] &= \text{Var}[T_s] + \mathbb{E}^2[T_s] \\ &\overset{1}{=} 1 - \mu(\mathbb{E}[T_s] - \mu) - (E[T_s] - \mu)^2 + \mathbb{E}^2[T_s] \\ &= 1 - \mu E[T_s] + \cancel{\mu^2} - \cancel{E^2[T_s]} + 2\mu E[T_s] - \cancel{\mu^2} + \cancel{E^2[T_s]} \\ &= 1 + \mu E[T_s] \end{aligned} \tag{D.1.10}$$

where (1) holds by (D.1.5) and (D.1.6). △

D.2 Models

D.2.1 CB-Probit Model

A Bayesian CB-Probit model is a categorical GLM which generates smulti-class outcomes $y_i \in \{1, ..., K\}$, $i = 1, ..., N$ by

$$\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{D.2.1a}$$

$$p_{ik} = \text{any CB-Probit category probabilities} \tag{D.2.1b}$$

$$y_i \sim \text{Cat}(p_{i1}, \dots p_{iK}). \tag{D.2.1c}$$

The form for the category probabilities in Eq. (D.2.1b) depends on the choice of CB model; for instance, for the CBM-Probit and CBC-Probit models we have

$$p_{ik}^{\text{CBM-Probit}} = \frac{\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)}{\sum_{\ell=1}^K \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_\ell)}$$

$$p_{ik}^{\text{CBC-Probit}} = \frac{\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)\prod_{h\neq k}\left(1-\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_h)\right)}{\sum_{\ell=1}^K \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_\ell)\prod_{h\neq \ell}\left(1-\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_h)\right)}$$

for standard Gaussian cdf Φ, known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$, and unknown parameters $\boldsymbol{B} \in \mathbb{R}^{M\times K}$ ($\boldsymbol{\beta}_k$ is used to designate the K-th column of $\boldsymbol{B}$).

D.2.2 IB-Probit Model

The base model for a CB-Probit model is an IB-Probit model. With a Gaussian prior, the model is:

$$\begin{aligned} \boldsymbol{\beta}_k &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad k = 1, ..., K \\ \widehat{y}_{ik} \mid \boldsymbol{\beta}_k &\overset{\text{ind}}{\sim} \text{Bernoulli}\big(\Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)\big), \quad i = 1, ..., N \end{aligned} \tag{D.2.2}$$

for known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\boldsymbol{B} \in \mathbb{R}^{M\times K}$ ($\boldsymbol{\beta}_k$ is used to designate the K-th column of $\boldsymbol{B}$). The binary responses $\widehat{y}_{ik}$ are the k-th element of $\widehat{\boldsymbol{y}}_i \in \{0,1\}^K$, where $\widehat{\boldsymbol{y}}_i = \boldsymbol{e}_{y_i}$ is the one-hot indicator vector with value of 1 only at entry $y_i \in \{1, \dots, K\}$.

D.2.3 Augmented IB-Probit Model

Following Albert & Chib (Albert & Chib, 1993), we may foster inference on the independent binary probit regression model by instead working with an augmented model.

$$\begin{aligned} \boldsymbol{\beta}_k &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad k = 1, ..., K \\ z_{ik} \mid \boldsymbol{\beta}_k &\overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T\boldsymbol{\beta}_k, 1), \quad i = 1, \dots, N \\ \widehat{y}_{ik} &= \begin{cases} 1 & z_{ik} \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, \dots, N \end{aligned} \tag{D.2.3}$$

where we have introduced augmented variables z_{ik}. We use $\boldsymbol{Z} \in \mathbb{R}^{N\times K}$ to represent the matrix whose (i,k)th entry is z_{ik}, and $\boldsymbol{z}_k$ to represent the kth column of $\boldsymbol{Z}$. As we will see in Sec. D.3.1, the augmented model is nice to work with, as it has exponential family complete conditionals.

D.3 Variational inference

Algorithm 2 provides closed-form coordinate ascent variational inference (CAVI) for the augmented IB-Probit model. By Eq. (3.2.2), this gives closed-form CAVI for any CB-Probit model.

D.3.1 Complete conditionals

The augmented IB-Probit model (Eq. (D.2.3)) contains Bayesian linear regression on the auxiliary variables $\boldsymbol{z}_k$. In this way, we obtain the complete conditionals

$$z_{ik} \mid \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K, \widehat{y}_{ik} \sim \begin{cases} \mathcal{N}_+\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_k\ ,\ 1\right), & \widehat{y}_{ik} = 1 \\ \mathcal{N}_-\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_k\ ,\ 1\right), & \text{otherwise} \end{cases} \tag{D.3.1}$$

Algorithm 2: Independent Binary Coordinate Ascent Variational Inference (IB-CAVI) for CB-Probit models

Input:

$\boldsymbol{y} \in \{1, \ldots, K\}^N$: N responses from K categories

$\boldsymbol{X} \in \mathbb{R}^{N \times M}$: Matrix whose ith row ($i = 1, \ldots, N$) gives the covariates associated with response y_i.

Output:

$\{(\widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)\}_{k=1}^K$: Parameters for $q(\boldsymbol{B}) = \prod_k \mathcal{N}(\beta_k | \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)$, variational density† (Eq. (D.3.3)) on regression weights

$\{\{\widetilde{\eta}_{ik}\}_{k=1}^K\}_{i=1}^N$: Parameters for $q(\boldsymbol{z}) = \prod_i \prod_k \mathcal{TN}(z_{ik} | \widetilde{\eta}_{ik}, 1, \Upsilon_{ik})$, variational density† (Eq. (D.3.3)) on auxiliary variables

Hyperparameters / Settings:

$(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$: Mean and covariance of prior density† on weights, $\pi(\boldsymbol{B}) = \prod_k \mathcal{N}(\boldsymbol{\beta}_k | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

$\{\widetilde{\boldsymbol{\mu}}_k^{(0)}\}_{k=1}^K$: Initial variational mean on regression weights

`Termination condition` : e.g. number of iterations or convergence threshold on $\texttt{ELBO}_{\text{IB}-\text{Probit}}$

†: Here $\mathcal{N}$ and $\mathcal{TN}$ refer to *densities* rather than *measures*.

for $k \leftarrow 1$ **to** K **do**

  $\widetilde{\boldsymbol{\Sigma}}_k \leftarrow \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$ `// Set` $\widetilde{\boldsymbol{\Sigma}}_k$ `for` $q(\boldsymbol{\beta}_k | \cdot, \widetilde{\boldsymbol{\Sigma}}_k)$ `via Eq.` (D.3.5b)

  while *termination condition not satisfied* **do**

    for $i \leftarrow 1$ **to** N **do**

      $\widetilde{\eta}_{ik} \leftarrow \boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k$ `// Update` $q(z_{ik} | \widetilde{\eta}_{ik})$ `via Eq.` (D.3.7)

      $\mathbb{E}_q[z_{ik}] \leftarrow \begin{cases} \widetilde{\eta}_{ik} + \dfrac{\phi(-\widetilde{\eta}_{ik})}{1 - \Phi(-\widetilde{\eta}_{ik})}, & \widehat{y}_{ik} = 1 \\ \widetilde{\eta}_{ik} - \dfrac{\phi(-\widetilde{\eta}_{ik})}{\Phi(-\widetilde{\eta}_{ik})}, & \text{otherwise} \end{cases}$ `// Expectation computed via Eq.` (D.3.6)

    end

    $\widetilde{\boldsymbol{\mu}}_k \leftarrow \widetilde{\boldsymbol{\Sigma}}_k \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{X}^T \mathbb{E}_q[\boldsymbol{z}_k]\right)$ `// Update` $q(\boldsymbol{\beta}_k | \widetilde{\mu}_k, \cdot)$ `via Eq.` (D.3.5a)

    (Optional) Compute $\texttt{ELBO}_{\text{IB}-\text{Probit}}$ via (D.3.8)

  end

end

where $\mathcal{N}_+$ and $\mathcal{N}_-$ are truncated normal distributions defined in Section D.1.2, and

$$\begin{aligned}\boldsymbol{\beta}_k \mid \boldsymbol{z}_k &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\\ \boldsymbol{\mu}_k &= \boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{z}_k\right),\\ \boldsymbol{\Sigma}_k &= \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\end{aligned}$$

D.3.2 Variational Family

We take the mean-field variational family for the augmented IB-Probit model (Eq. (D.2.3)) to have density given by

$$\begin{aligned}q(\boldsymbol{B}, \boldsymbol{Z}) &\overset{(1)}{=} q(\boldsymbol{B})q(\boldsymbol{Z})\\ &\overset{(2)}{=} \prod_{k=1}^{K} \underbrace{q(\boldsymbol{\beta}_k)}_{\mathcal{N}(\widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)} \prod_{i=1}^{N} \underbrace{q(z_{ik})}_{\mathcal{TN}(\widetilde{\eta}_{ik}, 1, \Upsilon_{ik})}\end{aligned} \tag{D.3.3}$$

$$\text{where } \Upsilon_{ik} = \begin{cases}\mathbb{R}^+, & \widehat{y}_{ik} = 1\\ \mathbb{R}^-, & \widehat{y}_{ik} = 0\end{cases} \tag{D.3.4}$$

Equality (1) is by mean-field assumption. Equality (2) holds without any additional assumption. Since the complete conditionals are both exponential families, the optimal variational factors with respect to the lower bound $\texttt{ELBO}_{\text{IB}}$ are in the same exponential families, with natural parameters given by the variational expectation of the natural parameters of the corresponding complete conditionals (as in Eq. C.0.1).

D.3.3 Coordinate Ascent Updates

Here we derive the parameters for the updates, using the notation of Eq. (D.3.3).

Updates to $\{q(\boldsymbol{\beta}_k)\}_{k=1}^K$ Since the natural parameters of a multivariate Gaussian are the precision and precision-weighted mean, we reparametrize the surrogate complete conditional for each $\boldsymbol{\beta}_k$ in Eq. (D.3.2) before taking variational expectations of the parameters. Hence, the optimal update to each $q(\boldsymbol{\beta}_k \mid \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)$ with respect to the objective $\texttt{ELBO}_{\text{IB}}$ is given by

$$\begin{aligned}\widetilde{\boldsymbol{\Sigma}}_k^{-1} &= \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}\left[\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X}\right] = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X}\\ \widetilde{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{\mu}}_k &= \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}\left[\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{z}_k\right]\\ &= \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\mathbb{E}_{q_{\boldsymbol{z}_k}}[\boldsymbol{z}_k]\end{aligned}$$

where $\mathbb{E}_q[\boldsymbol{z}_k]$ is given explicitly below.

Thus, in standard parameterization, we update

$$\widetilde{\boldsymbol{\mu}}_k = \widetilde{\boldsymbol{\Sigma}}_k\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\mathbb{E}_{q_{\boldsymbol{z}_k}}[\boldsymbol{z}_k]\right) \tag{D.3.5a}$$

$$\widetilde{\boldsymbol{\Sigma}}_k = \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \tag{D.3.5b}$$

where $\mathbb{E}_q[\boldsymbol{z}_k] \in \mathbb{R}^N$ has i-th entry given by

$$\mathbb{E}_q[z_{ik}] = \begin{cases}\widetilde{\eta}_{ik} + \dfrac{\phi(-\widetilde{\eta}_{ik})}{1 - \Phi(-\widetilde{\eta}_{ik})}, & \widehat{y}_{ik} = 1\\ \widetilde{\eta}_{ik} - \dfrac{\phi(-\widetilde{\eta}_{ik})}{\Phi(-\widetilde{\eta}_{ik})}, & \text{otherwise}\end{cases} \tag{D.3.6}$$

by properties of the truncated normal distribution (Section D.1.2). Recall that ϕ and Φ refer to the pdf and cdf, respectively, of the standard normal.

Updates to $\{q(z_{ik})\}_{ik}$ In (D.3.1), we saw that the surrogate complete conditional for each z_{ik} has the form $\mathcal{TN}(\eta_{ik}, 1, \Upsilon_{ik})$, where Υ_{ik} is defined as in (D.3.4). But since each such distribution is in the exponential family with natural parameter η_{ik}, the optimal update for each $q(z_{ik})$ is given by

$$\widetilde{\eta}_{ik} = \mathbb{E}[\boldsymbol{x}_i^T\boldsymbol{\beta}_k] = \boldsymbol{x}_i^T\widetilde{\boldsymbol{\mu}}_k \tag{D.3.7}$$

D.3.4 The Evidence Lower Bound

We provide the evidence lower bound for the IB-Probit model, $\texttt{ELBO}_{\text{IB}}$, in the case of independent $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ priors on $\boldsymbol{\beta}_k$ for $k = 1, ..., K$. Using $\widehat{\boldsymbol{Y}} \in \{0, 1\}^{N\times K}$ to represent the matrix whose ith row is the one-hot encoded vector $\widehat{\boldsymbol{y}}_i = \boldsymbol{e}(y_i)$, we have

$$\begin{aligned}\texttt{ELBO}(q) &= \underbrace{\mathbb{E}_q[\log p(\widehat{\boldsymbol{Y}}, \boldsymbol{Z}, \boldsymbol{\beta})]}_{\text{energy}} + \underbrace{-\mathbb{E}_q[\log q(\boldsymbol{X}, \boldsymbol{B})]}_{\text{entropy}}\\ &= \sum_{k=1}^{K}\Bigg[\underbrace{\sum_{i=1}^{N}\mathbb{E}_q[\log p(\widehat{y}_{ik}, z_{ik} \mid \boldsymbol{\beta}_k)]}_{\text{(A)}} + \underbrace{\mathbb{E}_q[\log p(\boldsymbol{\beta}_k)]}_{\text{(B)}} +\\ &\quad \underbrace{-\sum_{i=1}^{N}\mathbb{E}_q[\log q(z_{ik})]}_{\text{(C)}} + \underbrace{-\mathbb{E}_q \log[q(\boldsymbol{\beta}_k)]}_{\text{(D)}}\Bigg]\end{aligned} \tag{D.3.8}$$

Term (A) is given by a sum whose ith summand is

$$\begin{aligned}&\mathbb{E}_q[\log p(\widehat{y}_{ik}, z_{ik} \mid \boldsymbol{\beta}_k)]\\ &= \mathbb{E}_q\left[\log\left\{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(z_{ik} - \boldsymbol{x}_i^T\boldsymbol{\beta}_k)^2\right)\left(1_{z_{ik}<0}^{1_{\widehat{y}_{ik}=0}} 1_{z_{ik}\geq 0}^{1_{\widehat{y}_{ik}=1}}\right)\right\}\right]\\ &= -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\mathbb{E}(z_{ik} - \boldsymbol{x}_i^T\boldsymbol{\beta}_k)^2\\ &\qquad \cancel{-\tfrac{1}{2}\mathbb{E}_q\left[1_{\widehat{y}_{ik}=0}\log 1_{z_{ik}<0} + 1_{\widehat{y}_{ik}=1}\log 1_{z_{ik}\geq 0}\right]}\\ &= -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\mathbb{E}_q[z_{ik}^2] + \mathbb{E}_q[z_{ik}]\boldsymbol{x}_i^T\mathbb{E}_q[\boldsymbol{\beta}_k] - \tfrac{1}{2}\mathbb{E}_q[(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)^2]\\ &\overset{1}{=} -\tfrac{1}{2}(\log 2\pi + 1) - \tfrac{1}{2}\widetilde{\eta}_{ik}\mathbb{E}_q[z_{ik}] + \mathbb{E}_q[z_{ik}]\widetilde{\eta}_{ik} - \tfrac{1}{2}\mathbb{E}_q[(\boldsymbol{x}_i^T\boldsymbol{\beta}_k)^2]\\ &\overset{2}{=} -\tfrac{1}{2}(\log 2\pi + 1) + \tfrac{1}{2}\mathbb{E}_q[z_{ik}]\widetilde{\eta}_{ik} - \tfrac{1}{2}\left(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\Sigma}}_k\boldsymbol{x}_i + \widetilde{\eta}_{ik}^2\right)\\ &\overset{3}{=} -\tfrac{1}{2}(\log 2\pi + 1) + \tfrac{1}{2}\widetilde{\eta}_{ik}\delta_{\widehat{y}_{ik}}(\widetilde{\eta}_{ik}) - \tfrac{1}{2}\boldsymbol{x}_i^T\widetilde{\boldsymbol{\Sigma}}_k\boldsymbol{x}_i\end{aligned}$$

where

$$\delta_{\widehat{y}_{ik}}(\widetilde{\eta}_{ik}) := \begin{cases}\dfrac{\phi(-\widetilde{\eta}_{ik})}{1 - \Phi(-\widetilde{\eta}_{ik})}, & \widehat{y}_{ik} = 1\\ -\dfrac{\phi(-\widetilde{\eta}_{ik})}{\Phi(-\widetilde{\eta}_{ik})}, & \widehat{y}_{ik} = 0\end{cases}$$

Equation (1) holds by Eq. (D.3.7) and by Eq. (D.1.10), (2) holds by applying the second moment decomposition $\mathbb{E}_q[W^2] = \text{Var}_q[W] + \mathbb{E}_q^2[W]$ in the case where $W = \boldsymbol{x}_i^T\boldsymbol{\beta}_k$, and (3) holds by applying Eq. (D.1.9) to express

$\mathbb{E}_q[z_{ik}]$ as a perturbation of $\widetilde{\eta}_{ik}$. Recall that ϕ and Φ are the pdf and cdf, respectively, of the standard normal distribution.

Term (B) is the negative cross-entropy of two Gaussians. For instance, in the special case of a $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ prior, we have

$$\begin{aligned}\mathbb{E}_q[\log p(\boldsymbol{\beta}_k)] &= -\frac{M}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}_q[\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k] \\ &\overset{1}{=} -\frac{M}{2}\log(2\pi) - \frac{1}{2}\left(\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_k) + \widetilde{\boldsymbol{\mu}}_k^T\widetilde{\boldsymbol{\mu}}_k\right)\end{aligned}$$

where (1) holds since $\mathbb{E}_q[\boldsymbol{\beta}_k^T\boldsymbol{\beta}_k] = \sum_{m=1}^{M}\mathbb{E}_q[\beta_{km}^2] = \sum_{m=1}^{M}\mathrm{Var}_q[\beta_{km}] + \mathbb{E}_q^2[\beta_{km}]$.

Term (C) is the sum of entropies of truncated normal distributions, where the i-th element in the sum is

- the entropy of $\mathcal{N}_+(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\mu}}_k, 1)$ when $\widehat{y}_{ik} = 1$, in which case the entropy is given by Eq. (D.1.7)
- the entropy of $\mathcal{N}_-(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\mu}}_k, 1)$ when $\widehat{y}_{ik} = 0$, in which case the entropy is given by Eq. (D.1.8).

Term (D) is the entropy of the multivariate Gaussian $\mathcal{N}(\widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)$, which is given by

$$-\mathbb{E}_q\log[q(\boldsymbol{\beta}_k)] = \frac{1}{2}\ln|\widetilde{\boldsymbol{\Sigma}}_k| + \frac{M}{2}(1 + \ln 2\pi)$$

D.4 Computational complexity

The inference requires a one-time up-front computation of complexity $\mathcal{O}(M^3 + M^2N)$ due to the inversion in step Eq. (D.3.5b). Note that often the matrix $\boldsymbol{X}$ will be sparse, in which case the complexity of this up-front step can be reduced. Note that this up-front inversion could be avoided (at the cost of losing information about correlations across the M covariates) by tweaking the variational family in Eq. (D.3.3) to make a stronger (*fully-mean field*) variational assumption $q(\boldsymbol{B}) = \prod_{m=1}^{M}\prod_{k=1}^{K} q(\beta_{mk})$, where each $q(\beta_{mk})$ is the density of a univariate Gaussian $\mathcal{N}(\widetilde{\mu}_{mk}, \widetilde{\sigma}_{mk}^2)$. Inference would proceed identically as before, except that the variational covariance for the kth category $\widetilde{\boldsymbol{\Sigma}}_k$ would become a *diagonal* $M \times M$ covariance matrix whose mth entry is given by $\widetilde{\sigma}_{mk}^2 = [(\boldsymbol{\Sigma}_0)_{mk} + \sum_{i=1}^{N} x_{im}^2]^{-1}$. With this simplification, the up-front computation would have complexity $\mathcal{O}(MNK)$.

Afterwards, the computational complexity for a single CAVI update is $\mathcal{O}(MNK)$, where M is the number of covariates, N is the number of samples, and K is the number of categories. The complexity for each substep of a single CAVI update is given in the Table D.1 .

Moreover, the entire inference procedure (across all iterations) is embarassingly parallel over the K categories, so distributed computation can reduce the complexity for the all CAVI steps to $\mathcal{O}(MNI)$, where I is the number of CAVI iterations.

Variable	Step	Per-iteration complexity	Note
$\boldsymbol{B}$	covariance (D.3.5b)	pre-multiplied	$\underset{MxM}{\widetilde{\boldsymbol{\Sigma}}_k} = \underset{MxM}{\boldsymbol{\Sigma}}$ is the same for all k and constant over iterations.
$\boldsymbol{B}$	mean (D.3.5a)	$\mathcal{O}(MNK)$	$\underset{MxK}{\widetilde{\boldsymbol{\mu}}} = \underset{MxM}{\boldsymbol{\Sigma}}\underset{MxN}{\boldsymbol{X}^T}\underset{NxK}{\mathbb{E}_q[\boldsymbol{Z}]}$ the first two terms can be pre-multiplied
$\boldsymbol{Z}$	(D.3.7)	$\mathcal{O}(MNK)$	$\underset{NxK}{\widetilde{\boldsymbol{\eta}}} = \underset{NxM}{\boldsymbol{X}}\underset{MxK}{\widetilde{\boldsymbol{\mu}}}$
$\boldsymbol{Z}$	(D.3.6)	$\mathcal{O}(NK)$	We suppress the complexity of evaluating the Gaussian cdf.

Table D.1: *The computational complexity of CAVI updates for* IB-*Probit.* $\boldsymbol{B}$ is the matrix of regression weights and $\boldsymbol{Z}$ are auxiliary variables added for conditional conjugacy.

D.5 Sparsity considerations

When $N \times K$ is large, the matrix $\widetilde{\boldsymbol{\eta}}$ may not fit into memory. However, when the design matrix $\boldsymbol{X}$ is sparse , $\widetilde{\boldsymbol{\eta}}$ may be highly sparse ; indeed, $\widetilde{\eta}_{ik} = 0$ whenever at least one of $\{\widetilde{\boldsymbol{\mu}}_{mk}, \boldsymbol{X}_{im}\}$ is 0 for all $m = 1, ..., M$. In this setting, we can represent $\mathbb{E}_q[\boldsymbol{Z}]$ efficiently, since we can see from Eq. (D.3.6) that only two values are possible when $\eta_{ik} = 0$.

$$\mathbb{E}_q[z_{ik}] = \begin{cases} 2\phi(0), & \eta_{ik} = 0, y_i = k \\ -2\phi(0), & \eta_{ik} = 0, y_i \neq k \end{cases} \tag{D.5.1}$$

Define

$$\mathbb{E}_q[\boldsymbol{Z}]^* : (\mathbb{E}_q[\boldsymbol{Z}]^*)_{ik} = \begin{cases} (\mathbb{E}_q[\boldsymbol{Z}])_{ik}, & \eta_{ik} \neq 0 \\ 0, & \eta_{ik} = 0 \end{cases}$$

$$\mathbb{E}_q[\boldsymbol{Z}]^\dagger : (\mathbb{E}_q[\boldsymbol{Z}]^\dagger)_{ik} = \begin{cases} 1, & \eta_{ik} = 0, y_i = k \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}_q[\boldsymbol{Z}]^\ddagger : (\mathbb{E}_q[\boldsymbol{Z}]^\ddagger)_{ik} = \begin{cases} 1, & \eta_{ik} = 0, y_i \neq k \\ 0, & \text{otherwise} \end{cases}$$

Then we can avoid representing $\mathbb{E}_q[\boldsymbol{Z}]$ as a large dense $N \times K$ matrix of floats by rewriting Eq. (D.3.6) in matrix form as

$$E_q[\boldsymbol{Z}] = \mathbb{E}_q[\boldsymbol{Z}]^* + 2\phi(0)\left(\mathbb{E}_q[\boldsymbol{Z}]^\dagger - \mathbb{E}_q[\boldsymbol{Z}]^\ddagger\right)$$

E Variational inference for CB-Logit Models

Here we present closed-form variational inference for CB-Logit models. The inference follows naturally from our IB-CAVI procedure in Algorithm 1.

E.1 Distributional preliminaries

Definition E.1.1. A non-negative random variable X has a *Pòlya-Gamma distribution* (Polson et al., 2013) with pa-

rameters $b > 0$ and $c \in \mathbb{R}$, denoted as $X \sim \text{PG}(b, c)$, if

$$X \overset{D}{=} \frac{1}{2\pi^2} \sum_{r=1}^{\infty} \frac{\gamma_r}{(r-1/2)^2 + c^2/(4\pi^2)}$$

where the $\gamma_r \sim \text{Gamma}(b, 1)$ are independent Gamma random variables, and where $\overset{D}{=}$ indicates equality in distribution. $\triangle$

The density of a $\text{PG}(b, c)$ random variable can be written as (Polson et al., 2013):

$$f(x; b, c) = \cosh^b(c/2) e^{-\frac{c^2}{2}x} g(x; b, 0) \tag{E.1.1}$$

where $g(x; b, 0)$ is the density of a $\text{PG}(b, 0)$ random variable

$$g(x; b, 0) = \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)}{\Gamma(n+1)} \frac{(2n+b)}{\sqrt{2\pi x^3}} e^{-\frac{(2n+b)^2}{8x}}$$

So f is constructed from g via an *exponential tilt* and a renormalization.

The expectation of a Pòlya-Gamma random variable is given by (Polson et al., 2013) as

$$\mathbb{E}[X] = \frac{b}{2c} \tanh(c/2) = \frac{b}{2c} \frac{e^c - 1}{e^c + 1}$$

E.2 Models

E.2.1 CB-Logit Model

A Bayesian CB-Logit model is a categorical GLM which generates multi-class outcomes $y_i \in \{1, ..., K\}$, $i = 1, ..., N$ by

$$\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \tag{E.2.1a}$$
$$p_{ik} = \text{any CB-Logit category probabilities} \tag{E.2.1b}$$
$$y_i \sim \text{Cat}(p_{i1}, \dots p_{iK}). \tag{E.2.1c}$$

The form for the category probabilities in Eq. (E.2.1b) depends on the choice of CB model; for instance, for the CBM-Logit and CBC-Logit models we have

$$p_{ik}^{\text{CBM-Logit}} = \frac{L(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)}{\sum_{\ell=1}^K L(\boldsymbol{x}_i^T \boldsymbol{\beta}_\ell)}$$

$$p_{ik}^{\text{CBC-Logit}} = \frac{L(\boldsymbol{x}_i^T \boldsymbol{\beta}_k) \prod_{h \neq k} \left(1 - L(\boldsymbol{x}_i^T \boldsymbol{\beta}_h)\right)}{\sum_{\ell=1}^K L(\boldsymbol{x}_i^T \boldsymbol{\beta}_\ell) \prod_{h \neq \ell} \left(1 - L(\boldsymbol{x}_i^T \boldsymbol{\beta}_h)\right)}$$

for standard Logistic cdf L, known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$, and unknown parameters $\boldsymbol{B} \in \mathbb{R}^{M \times K}$ ($\boldsymbol{\beta}_k$ is used to designate the k-th column of $\boldsymbol{B}$).

E.2.2 IB-Logit Model

The base model for a CB-Logit model is an IB-Logit model. With a Gaussian prior, the model is:

$$\begin{aligned} \boldsymbol{\beta}_k &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad k = 1, ..., K \\ \widehat{y}_{ik} \mid \boldsymbol{\beta}_k &\overset{\text{ind}}{\sim} \text{Bernoulli}\big(L(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)\big), \quad i = 1, ..., N \end{aligned} \tag{E.2.2}$$

for known binary responses $\widehat{y}_{ik}$, known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\boldsymbol{B} \in \mathbb{R}^{M \times K}$ ($\boldsymbol{\beta}_k$ is used to designate the K-th column of $\boldsymbol{B}$). We write $\widehat{\boldsymbol{Y}} \in \{0, 1\}^{N \times K}$ to represent the matrix with one-hot encoded rows such that $\widehat{\boldsymbol{Y}}_{i,k} = 1$ if the ith outcome was the kth category (i.e. if $y_i = k$), and $\widehat{\boldsymbol{y}}_k$ to represent the kth column of $\widehat{\boldsymbol{Y}}$.

E.2.3 Augmented IB-Logit Model

The main idea is to introduce auxiliary latent variables $\boldsymbol{\omega}_k = (\omega_{1k}, ..., \omega_{Nk})$ with Polya-Gamma distribution to make the model of Eq. (E.2.2) fully conditionally conjugate. The model is fully conditionally conjugate in the sense that the complete conditionals and the priors form conjugate pairs; that is $p(\boldsymbol{\beta}_k \mid \boldsymbol{w}_k, \widehat{\boldsymbol{y}}_k)$ is in the same family (Gaussian) as $p(\boldsymbol{\beta}_k)$, and each $p(w_{ik} \mid \boldsymbol{\beta}_k, \widehat{\boldsymbol{y}}_k)$ is in the same family (PG) as $p(w_{ik})$. Thus, inference on the augmented model is easy. Marginalizing over these auxillary variables in the posterior distribution yields the desired target posterior on $\boldsymbol{B} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$. We use $\boldsymbol{\Omega} \in \mathbb{R}^{N \times K}$ to represent the matrix whose kth column is $\boldsymbol{\omega}_k$.

We now form the augmented model. Conditional on each $\boldsymbol{\beta}_k$, we take $\{(\widehat{y}_{ik}, \omega_{ik})\}_{i=1}^N$ to be independent random pairs such that $\widehat{y}_{ik}$ and ω_{ik} are also independent, where

$$\begin{aligned} \boldsymbol{\beta}_k &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad k = 1, ..., K \\ \widehat{y}_{ik} \mid \boldsymbol{\beta}_k &\overset{\text{ind}}{\sim} \text{Bernoulli}\left(\frac{\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}_k\}}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}_k\}}\right), \quad i = 1, ..., N \\ \omega_{ik} \mid \boldsymbol{\beta} &\overset{\text{ind}}{\sim} \text{PG}(1, \boldsymbol{x}_i^T \boldsymbol{\beta}_k) \end{aligned} \tag{E.2.3}$$

The *augmented posterior density* for the kth binary logistic regression is given by

$$p(\boldsymbol{\beta}_k, \boldsymbol{\omega}_k \mid \widehat{\boldsymbol{y}}_k) \propto \left[\prod_{i=1}^N p(\widehat{y}_{ik} \mid \boldsymbol{\beta}_k) p(\omega_{ik} \mid \boldsymbol{\beta}_k)\right] p(\boldsymbol{\beta}_k)$$

And clearly

$$\int_{\mathbb{R}_+^N} p(\boldsymbol{\beta}_k, \boldsymbol{\omega}_k \mid \widehat{\boldsymbol{y}}_k) d\boldsymbol{\omega}_k = p(\boldsymbol{\beta}_k \mid \widehat{\boldsymbol{y}}_k)$$

which is the target posterior density for the k binary logistic regression.

A straightforward argument (see Section F.3.1) reveals that the complete conditionals for the kth binary logistic regression are given by

$$(\omega_{ik} \mid \boldsymbol{\beta}_k) \sim \text{PG}(1, \boldsymbol{x}_i^T \boldsymbol{\beta}_k) \tag{E.2.4a}$$
$$(\boldsymbol{\beta}_k \mid \widehat{\boldsymbol{y}}_k, \boldsymbol{\omega}_k) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}_k}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}_k}) \tag{E.2.4b}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}_k} = \left(\boldsymbol{X}^T \boldsymbol{W}_{\boldsymbol{\omega}_k} \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \tag{E.2.5}$$

$$\boldsymbol{\mu}_{\boldsymbol{\omega}_k} = \boldsymbol{\Sigma}_{\boldsymbol{\omega}} \left(\boldsymbol{X}^T \boldsymbol{\kappa}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right) \tag{E.2.6}$$

and

$$\boldsymbol{\kappa}_k = (\widehat{y}_{1k} - \frac{1}{2}, ..., \widehat{y}_{Nk} - \frac{1}{2})^T \tag{E.2.7a}$$

$$\boldsymbol{W}_{\boldsymbol{\omega}_k} \text{ is the diagonal matrix of } \omega_{ik}\text{'s} \tag{E.2.7b}$$

Another straightforward argument (see Section 2 of (Choi et al., 2013)) reveals that the complete conditionals (E.2.4) for the Bayesian logistic regression model under Pòlya-Gamma augmentation form a valid Gibbs sampler.

E.3 Variational inference

We are able to easily construct a mean-field variational inference algorithm using these complete conditionals (E.2.4), since the complete conditionals are in the exponential family. Algorithm 3 provides closed-form coordinate ascent variational inference (CAVI) for the augmented IB-Logit model. By Eq. (3.2.2), this gives closed-form CAVI for any CB-Logit model.

Proposition E.3.1. *An optimal mean-field coordinate ascent variational inference algorithm for estimating the posterior of the* IB*-Logit model with Pòlya-Gamma augmentation (Eq.* (E.2.3)*) by using a member of the variational family whose density factorizes as*

$$q(\boldsymbol{\Omega}, \boldsymbol{B}) = q(\boldsymbol{\Omega})q(\boldsymbol{B}) \tag{E.3.1}$$

can be obtained by taking the variational family to have the further factorization

$$q(\boldsymbol{\Omega}, \boldsymbol{B}) = \prod_{k=1}^{K} \underbrace{q(\boldsymbol{\beta}_k)}_{\mathcal{N}(\boldsymbol{\beta}_k; \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)} \prod_{i=1}^{N} \underbrace{q(\omega_{ik})}_{PG(\omega_{ik}; \widetilde{b_{ik}}, \widetilde{c_{ik}})} \tag{E.3.2}$$

with parameter updates given by

$$\widetilde{b_{ik}} = 1 \tag{E.3.3a}$$

$$\widetilde{c_{ik}} = \left(\boldsymbol{x}_i^T \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{x}_i + (\boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k)^2 \right)^{1/2} \tag{E.3.3b}$$

$$\widetilde{\boldsymbol{\Sigma}}_k = \left(\boldsymbol{X}^T \boldsymbol{W}_{\mathbb{E}_q[\boldsymbol{\omega}_k]} \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1} \tag{E.3.4a}$$

$$\widetilde{\boldsymbol{\mu}}_k = \widetilde{\boldsymbol{\Sigma}}_k \left(\boldsymbol{X}^T \boldsymbol{\kappa}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \tag{E.3.4b}$$

where $\boldsymbol{W}_{\mathbb{E}_q[\boldsymbol{\omega}_k]}$ *is the* $N \times K$ *diagonal matrix with diagonal entries*

$$\mathbb{E}_q[\omega_{ik}] = \frac{\widetilde{b_{ik}}}{2\widetilde{c_{ik}}} \frac{e^{\widetilde{c_{ik}}} - 1}{e^{\widetilde{c_{ik}}} + 1} \tag{E.3.5}$$

and where $\boldsymbol{\kappa}_k$ *was defined in Eq.* (E.2.7a).

Proof. The complete conditionals (E.2.4) are in the exponential family. For the Gaussian this is well-known. For the $\text{PG}(1, c_i)$ distribution, this is immediate from Eq. (F.3.1). Due to the membership of the complete conditionals in the exponential family, we can can apply Eq. (C.0.1) to determine that the optimal variational updates are in the same exponential family, with parameters given below. The additional independence structure in the variational family is obtained without further approximation by application of a known recipe (Blei et al., 2017).

Updating the variational distribution on $\boldsymbol{\beta}$. By Eq. (C.0.1), the optimal variational distribution at any update is Normal. The natural parameters of $\mathcal{N}(\boldsymbol{a}, \boldsymbol{B})$ are given by

$$\widetilde{\boldsymbol{\eta}}(\boldsymbol{a}, \boldsymbol{B}) = (\boldsymbol{B}^{-1}, \boldsymbol{B}^{-1}\boldsymbol{a}) := (\widetilde{\boldsymbol{\eta}}_1^{\mathcal{N}}, \widetilde{\boldsymbol{\eta}}_2^{\mathcal{N}}). \tag{E.3.6}$$

For the variational normal distribution on $\boldsymbol{\beta}_k$, we find that the the first coordinate of the natural variational parameter is given by:

$$\begin{aligned}
\widetilde{\boldsymbol{\eta}}_1^{\mathcal{N}} &= \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}[\boldsymbol{\eta}_1^{\mathcal{N}}] \overset{\text{(E.3.6)}}{=} \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}[\boldsymbol{\Sigma}_{\boldsymbol{\omega}_k}^{-1}] \\
&\overset{\text{(E.2.5)}}{=} \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}[\boldsymbol{X}^T \boldsymbol{W}_{\boldsymbol{\omega}_k} \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}] \\
&= \boldsymbol{X}^T \boldsymbol{W}_{\mathbb{E}_q[\boldsymbol{\omega}_k]} \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}
\end{aligned}$$

and therefore, by inverting the natural parameter transformation (E.3.6)

$$\widetilde{\boldsymbol{\Sigma}}_k = (\widetilde{\boldsymbol{\eta}}_1^{\mathcal{N}})^{-1} = \left(\boldsymbol{X}^T \boldsymbol{W}_{\mathbb{E}_q[\boldsymbol{\omega}_k]} \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1}$$

Similarly, the second coordinate of the natural variational parameter is given by

$$\begin{aligned}
\widetilde{\boldsymbol{\eta}}_2^{\mathcal{N}} &= \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}[\boldsymbol{\eta}_2^{\mathcal{N}}] = \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}[\boldsymbol{\Sigma}_{\boldsymbol{\omega}_k}^{-1} \boldsymbol{\mu}_{\boldsymbol{\omega}_k}] \\
&\overset{\text{(E.2.6)}}{=} \mathbb{E}_{q_{-\boldsymbol{\beta}_k}}[\boldsymbol{X}^T \boldsymbol{\kappa}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0] \\
&= \boldsymbol{X}^T \boldsymbol{\kappa}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0
\end{aligned}$$

and therefore, by inverting the natural parameter transformation (E.3.6)

$$\widetilde{\boldsymbol{\mu}}_k = (\widetilde{\boldsymbol{\eta}}_1^{\mathcal{N}})^{-1} \widetilde{\boldsymbol{\eta}}_2^{\mathcal{N}} = \widetilde{\boldsymbol{\Sigma}}_k \left(\boldsymbol{X}^T \boldsymbol{\kappa}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right)$$

Updating the variational distribution on $\boldsymbol{\Omega}$. By (E.2.4a), the complete conditional on each ω_{ik} has a PG distribution. Moreover,

$$\boldsymbol{\eta}_{ik}^{\text{PG}}(c_{ik}) = c_{ik}^2 \tag{E.3.7}$$

is a natural parameter for the $\text{PG}(1, c_{ik})$ distribution, as is immediate from (F.3.1).

Thus, we apply Eq. (C.0.1) to determine that the optimal variational update is also PG with natural parameter given by

Algorithm 3: Independent Binary Coordinate Ascent Variational Inference (IB-CAVI) for CB-Logit models

Input:

$\boldsymbol{y} \in \{1, \ldots, K\}^N$: A vector of N conditionally independent responses from K categories

$\boldsymbol{X} \in \mathbb{R}^{N \times M}$: A matrix whose ith row ($i = 1, \ldots, N$) gives the covariates associated with response y_i.

Output:

$\{(\widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)\}_{k=1}^K$: Parameters for $q(\boldsymbol{B}) = \prod_k \mathcal{N}(\boldsymbol{\beta}_k \mid \widetilde{\boldsymbol{\mu}}_k, \widetilde{\boldsymbol{\Sigma}}_k)$, variational density† (Eq. (E.3.2)) on regression weights

$\{\{\widetilde{c}_{ik}\}_{k=1}^K\}_{i=1}^N$: Parameters for $q(\boldsymbol{\Omega}) = \prod_i \prod_k \text{PG}(\omega_{ik} \mid 1, \widetilde{c_{ik}})$, variational density† (Eq. (E.3.2)) on auxiliary variables

Hyperparameters / Settings:

$(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$: Mean and covariance of prior density† on weights, $\pi(\boldsymbol{B}) = \prod_k \mathcal{N}(\boldsymbol{\beta}_k | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

$\{(\widetilde{\boldsymbol{\mu}}_k^{(0)}, \widetilde{\boldsymbol{\Sigma}}_k^{(0)})\}_{k=1}^K$: Initial variational parameters on regression weights

`Termination condition:` e.g. number of iterations, or convergence threshold on $\texttt{ELBO}_{\text{IB}-\text{Logit}}$

†: Here $\mathcal{N}$ and PG refer to *densities* rather than *measures*.

for $k \leftarrow 1$ **to** K **do**

  $\boldsymbol{\kappa}_k \leftarrow (\widehat{y}_{1k} - \frac{1}{2}, ..., \widehat{y}_{Nk} - \frac{1}{2})^T$; `// where` $\widehat{y}_{ik} := 1$ `iff` $y_i = k$ `(Sec. E.2.2)`

  while *termination condition not satisfied* **do**

    for $i \leftarrow 1$ **to** N **do**

      $\widetilde{c_{ik}} \leftarrow \left(\boldsymbol{x}_i^T \widetilde{\boldsymbol{\Sigma}}_k \boldsymbol{x}_i + (\boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k)^2 \right)^{1/2}$; `// Update` $q(\omega_{ik} \mid 1, \widetilde{c_{ik}})$ `via Eq.` (E.3.3b)

      $\mathbb{E}_q[\omega_{ik}] \leftarrow \frac{1}{2\widetilde{c_{ik}}} \frac{e^{\widetilde{c_{ik}}} - 1}{e^{\widetilde{c_{ik}}} + 1}$; `// Expectation computed via Eq.` (E.3.5)

    end

    $\boldsymbol{W}_k \leftarrow$ diag. matrix from $(\mathbb{E}_q[\omega_{ik}])_{i=1}^N$;

    $\widetilde{\boldsymbol{\Sigma}}_k \leftarrow \left(\boldsymbol{X}^T \boldsymbol{W}_k \boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1}$; `// Update` $q(\boldsymbol{\beta}_k | \cdot, \widetilde{\boldsymbol{\Sigma}}_k)$ `via Eq.` (E.3.4a)

    $\widetilde{\boldsymbol{\mu}}_k \leftarrow \widetilde{\boldsymbol{\Sigma}}_k \left(\boldsymbol{X}^T \boldsymbol{\kappa}_k + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right)$; `// Update` $q(\boldsymbol{\beta}_k | \widetilde{\boldsymbol{\mu}}_k, \cdot)$ `via Eq.` (E.3.4b)

    (Optional) Compute $\texttt{ELBO}_{\text{IB}-\text{Logit}}$ via (E.3.9)

  end

end

$$
\begin{aligned}
\widetilde{\boldsymbol{\eta}}_{ik}^{\text{PG}} &\overset{Eq.\ (C.0.1)}{=} \mathbb{E}_{q_{-\omega_{ik}}}[\boldsymbol{\eta}_{ik}^{\text{PG}}] \\
&\overset{(E.3.7)}{=} \mathbb{E}_{q_{-\omega_{ik}}}[c_{ik}^2] \\
&= \text{Var}_{q_{-\omega_{ik}}}[c_{ik}] + \mathbb{E}_{q_{-\omega_{ik}}}[c_{ik}]^2 \\
&\overset{(E.2.4a)}{=} \text{Var}_{q_{\boldsymbol{\beta}_k}}[\boldsymbol{x}_i^T\boldsymbol{\beta}_k] + \mathbb{E}_{q_{\boldsymbol{\beta}_k}}[\boldsymbol{x}_i^T\boldsymbol{\beta}_k]^2 \\
&= \boldsymbol{x}_i^T \text{Var}_{q_{\boldsymbol{\beta}_k}}[\boldsymbol{\beta}_k]\, \boldsymbol{x}_i + (\boldsymbol{x}_i^T \mathbb{E}_{q_{\boldsymbol{\beta}_k}}[\boldsymbol{\beta}_k])^2 \\
&= \boldsymbol{x}_i^T \widetilde{\boldsymbol{\Sigma}}_k\, \boldsymbol{x}_i + (\boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k)^2
\end{aligned}
$$

and therefore, by inverting the natural parameter transformation (E.3.7)

$$
\widetilde{c}_{ik} = \left(\boldsymbol{x}_i^T \widetilde{\boldsymbol{\Sigma}}_k\, \boldsymbol{x}_i + (\boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k)^2 \right)^{1/2} \tag{E.3.8}
$$

where it suffices to take the positive square root since the density of $\text{PG}(1, c)$ is symmetric around $c = 0$. □

E.3.1 The evidence lower bound

Here we provide an expression for the ELBO.

Proposition E.3.2. *The Evidence Lower Bound (ELBO) for the* IB-*Logit with Pòlya-Gamma augmentation (Eq. (E.2.3)) when using the mean-field variational approximation (E.3.1) is given by*

$$
\mathit{ELBO}[q(\boldsymbol{B}, \boldsymbol{\Omega})] = \sum_{k=1}^{K} \mathit{ELBO}[q(\boldsymbol{\beta}_k, \boldsymbol{\omega}_k)]
$$

where

$$
\begin{aligned}
\mathit{ELBO}[q(\boldsymbol{\beta}_k, \boldsymbol{\omega}_k)] &= \frac{1}{2}d + \frac{1}{2}\log|\widetilde{\boldsymbol{\Sigma}}_k| + \frac{1}{2}\log|\boldsymbol{\Sigma}_0^{-1}| \\
&- \frac{1}{2}(\widetilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\widetilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0) - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_0^{-1}\widetilde{\boldsymbol{\Sigma}}_k) \\
&+ \sum_{i=1}^{N} (\widehat{y}_{ik} - \frac{1}{2})\, \boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k - \log\left[1 + \exp(-\widetilde{c_{ik}})\right] - \frac{1}{2}\widetilde{c_{ik}}
\end{aligned} \tag{E.3.9}
$$

and where d is the number of rows of $\boldsymbol{B}$.

Proof.

$$
\begin{aligned}
&\texttt{ELBO}[q(\boldsymbol{B}, \boldsymbol{\Omega})] \\
&\quad = \mathbb{E}_{q(\boldsymbol{B},\boldsymbol{\Omega})}[\log p(\widehat{\boldsymbol{Y}}, \boldsymbol{B}, \boldsymbol{\Omega})] - \mathbb{E}_{q(\boldsymbol{B},\boldsymbol{\Omega})}[\log q(\boldsymbol{B}, \boldsymbol{\Omega})] \\
&\quad = \sum_{k=1}^{K} \Bigg[\mathbb{E}_{q(\boldsymbol{\beta}_k)}[\log p(\boldsymbol{\beta}_k)] \\
&\qquad\quad + \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta}_k)} \mathbb{E}_{q(\omega_{ik})}[\log p(\widehat{y}_{ik}, \omega_{ik} \mid \boldsymbol{\beta}_k)] \\
&\qquad\quad - \mathbb{E}_{q(\boldsymbol{\beta}_k)}[\log q(\boldsymbol{\beta}_k)] \\
&\qquad\quad - \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta}_k)} \mathbb{E}_{q(\omega_{ik})}[\log q(\omega_{ik})] \Bigg] \\
&\quad = \sum_{k=1}^{K} \Bigg[- \texttt{KL}\big(q_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}_k) \;||\; p_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}_k)\big) \\
&\qquad\quad + \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{\beta}_k)} \mathbb{E}_{q(\omega_{ik})}[\log p(\widehat{y}_{ik}, \omega_{ik} \mid \boldsymbol{\beta}_k) \\
&\qquad\quad - \log q(\omega_{ik})] \Bigg]
\end{aligned} \tag{E.3.10}
$$

The first term is the negative KL divergence between two Gaussians, which is well-known (and is given by the first line of Eq. (E.3.9)).

For the second term, we read Lemma 1 of (Durante & Rigon, 2019) from right to left to obtain

$$
\begin{aligned}
&\mathbb{E}_{q(\omega_{ik})}[\log p(\widehat{y}_{ik}, \omega_{ik} \mid \boldsymbol{\beta}_k) - \log q(\omega_{ik})] = \log \bar{p}(\widehat{y}_{ik} \mid \boldsymbol{\beta}_k) \\
&\quad = (\widehat{y}_{ik} - \frac{1}{2})\, \boldsymbol{x}_i^T \boldsymbol{\beta}_k - \frac{1}{2}\widetilde{c_{ik}} \\
&\qquad - \frac{1}{4}\widetilde{c_{ik}}^{-1} \tanh(\frac{1}{2}\widetilde{c_{ik}})\left[(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)^2 - \widetilde{c_{ik}}^2\right] \\
&\qquad - \log\left[1 + \exp(-\widetilde{c_{ik}})\right]
\end{aligned} \tag{E.3.11}
$$

where $\log \bar{p}(\widehat{y}_{ik} \mid \boldsymbol{\beta}_k) \leq \log p(\widehat{y}_{ik} \mid \boldsymbol{\beta}_k)$ is the well-known quadratic lower-bound given by (Jaakkola & Jordan, 2000).[4]

Taking the expectation w.r.t $q_{\boldsymbol{\beta}_k}$ of Eq. (E.3.11), we obtain

$$
\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\beta}_k)} \mathbb{E}_{q(\omega_{ik})}[\log p(\widehat{y}_{ik}, \omega_{ik} \mid \boldsymbol{\beta}_k) - \log q(\omega_{ik})] \\
&\quad = (\widehat{y}_{ik} - \frac{1}{2})\, \boldsymbol{x}_i^T \widetilde{\boldsymbol{\mu}}_k - \frac{1}{2}\widetilde{c_{ik}} - \log\left[1 + \exp(-\widetilde{c_{ik}})\right]
\end{aligned} \tag{E.3.12}
$$

where the third term in the sum disappears since $\mathbb{E}_{q(\boldsymbol{\beta}_k)}[(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)^2] = \widetilde{c_{ik}}^2$, as we saw in the argument leading to Eq. (E.3.8).

Taking the sum of Eq. (E.3.12) across N observations produces the second term in Eq. (E.3.10). □

[4] That is, the *exact* ELBO for the IB-Logit model after Pòlya-Gamma augmentation has a summand which can be expressed as the expected value of the the well-known quadratic lower-bound given by (Jaakkola & Jordan, 2000).

E.4 Computational complexity

The complexity for each iteration of CAVI for an IB-Logit model is $\mathcal{O}(M^3K + NM^2K)$. Details on each substep are given in Table E.1. Note in particular that, unlike with the IB-Probit model, the IB-Logit model requires a matrix inversion *at each step of inference*, rather than just once up-front. This increases the per-iteration complexity from $\mathcal{O}(MNK)$. (Recall that Sec. D.4 provides more information on the complexity of CAVI for IB-Probit.)

If one imposes an additional variational assumption beyond that given in Eq. E.3.1, namely that each category's regression weights are independent across covariates, $q(\boldsymbol{\beta}_k) = \prod_{m=1}^{M} q(\beta_{mk})$, then the additional computational complexity imposed by IB-Logit over IB-Probit can be avoided. This strategy may make sense when the choice of link (Logit over Probit) is more important than modeling the correlations in regression weights across covariates.

As with the IB-Probit model, the entire inference procedure (across all iterations) is embarassingly parallel over the K categories. Thus, distributed computation over K nodes can reduce the complexity for the entire inference procedure to $\mathcal{O}((M^3 + NM^2)I)$, where I is the number of CAVI iterations. Recall also that sparsity of the matrix $\boldsymbol{X}$ can reduce the complexity of these steps.

Variable	Step	Per-iteration complexity	Note
$\boldsymbol{B}$	covariance (E.3.4a)	$\mathcal{O}(M^3K + NM^2K)$	matrix inversion
$\boldsymbol{B}$	mean (E.3.4b)	$\mathcal{O}(MNK + M^2K)$	covariance matrix not pre-computable
$\boldsymbol{\Omega}$	(E.3.3)	$\mathcal{O}(NM^2 + NMK)$	$\underset{NxM}{\boldsymbol{X}}$ $\underset{MxM}{\tilde{\boldsymbol{\Sigma}}}$ $\underset{MxN}{\boldsymbol{X}^T}$ and $\underset{NxM}{\boldsymbol{X}}$ $\underset{MxK}{\tilde{\boldsymbol{\mu}}}$

Table E.1: *The computational complexity of CAVI updates for* IB-*Probit.* $\boldsymbol{B}$ is the matrix of regression weights and $\boldsymbol{\Omega}$ are auxiliary variables added for conditional conjugacy.

F Alternative methods for Bayesian inference in categorical GLMs

In this section, we provide further information about alternative approaches to Bayesian inference for categorical GLMs. For orientation, see Table 1, for which an extended caption is given in Sec. F.1. In particular, the first five rows of Table 1 provide alternative approaches to CB-Models with IB-CAVI.

- In Sec. F.2, we describe automatic differentation variational inference, which can be used for Bayesian inference with the softmax model (row 1). We include this approach in our experiments.
- We do not consider the MNP models (rows 2 and 5) due to the lack of closed-form category probabilities, which can complicate inference in high dimensions.
- In Sec. F.3, we provide the construction of a Gibbs sampler for the softmax (more specifically, for the multi-logit model, which is the softmax model but with one category's vector of regression weights fixed to $\mathbf{0}$ for identifiability) after Pòlya-Gamma augmentation (row 3). We include this approach in our experiments. We cannot construct closed-form CAVI for softmax after Pòlya-Gamma augmentation , as we show in Sec. F.4.
- In Sec. F.5, we consider the stick-breaking construction of the softmax (row 4). We highlight the category asymmetry of this method, which causes us to not consider this approach further in our experiments.

F.1 Extended caption for Table 1

An extended caption for Table 1 follows. See the main body of the text for citations for these methods.

Further details on columns:

- `Category symmetry` refers to symmetric handling of categories.
- `Latent linear regression` reports the existence of latent auxiliary variables z_i, one for which observation, for which the regression weights $\boldsymbol{\beta}$ have a linear regression likelihood. (This enables easy extensibility, e.g. to hierarchical models or variable selection priors.)
- `Auxiliary variable independence` is satisfied when the latent auxiliary variables, one for each observation, are conditionally independent across categories given all observations and all other unobserved random variables. (Non-existence of such auxiliary variables is considered to meet the criterion.)
- `Closed-form likelihood` refers to closed-form category probabilities in the marginal likelihood.
- `Conditional conjugacy` refers to the state whereby a (non-trivial) conjugate prior exists for each complete conditional.
- `Closed-form variational inference` refers to the existence of a known coordinate ascent variational inference algorithm with closed-form updates.
- `Embarassingly parallel across categories` refers to the state where the inference can be performed separately on each category's regression weights.

Further details on specific cells: The lack of closed-form CAVI for Softmax+PGA is reviewed in Sec. F.4. The category asymmetry of the SB-Softmax+PGA method is discussed in Sec. F.5. The latent linear regression property of IB-CAVI can be exploited for closed-form hierarchical modeling, as mentioned in Sec. C, but we do not consider such models in this paper.

F.2 Automatic differentiation variational inference

Automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017) is a generic variational inference method that applies to a large class of Bayesian models. Its objective function is known as the ADVI evidence lower bound (ADVI ELBO):

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\boldsymbol{I})}\left[\log p\left(\boldsymbol{y}, T^{-1}\left(S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})\right)\right) + \log\left|\det J_{T^{-1}}\left(S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})\right)\right|\right] + \mathbb{H}[q(\boldsymbol{\zeta};\boldsymbol{\lambda})] \quad \text{(F.2.1)}$$

where $\boldsymbol{\lambda}$ are the variational parameters, $T : \Theta \to \mathbb{R}^P$, $\boldsymbol{\theta} \mapsto \boldsymbol{\zeta}$ is a differentiable bijection to give the model parameters $\boldsymbol{\theta}$ unbounded support, and $S_{\boldsymbol{\lambda}} : \mathbb{R}^P \to \mathbb{R}^P$, $\boldsymbol{\zeta} \mapsto \boldsymbol{\epsilon}$ is a (deterministic) standardization function.

The gradients for the ADVI objective are given in (Kucukelbir et al., 2017). We assume a Gaussian mean-field variational family on $\boldsymbol{\zeta} \in \mathbb{R}^p$, the transformed unobserved random variables, i.e. the variational parameters are $\boldsymbol{\lambda} = \left(\widetilde{\boldsymbol{\mu}}, \text{diag}(\widetilde{\boldsymbol{\sigma}}^2)\right)$ and the variational density is given by

$$q(\boldsymbol{\zeta};\boldsymbol{\lambda}) = \prod_{p=1}^{P} \underbrace{q(\zeta_p;\boldsymbol{\lambda}_p)}_{\mathcal{N}(\widetilde{\mu}_p, \widetilde{\sigma}_p^2)}$$

Under this Gaussian mean field assumption, the gradients are given by (Kucukelbir et al., 2017)

$$\nabla_{\widetilde{\boldsymbol{\mu}}}\mathcal{L} = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\boldsymbol{I})}\left[\underbrace{\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\theta})}_{1\times p}\underbrace{\nabla_{\boldsymbol{\zeta}} T^{-1}(\boldsymbol{\zeta})}_{p\times p} + \underbrace{\nabla_{\boldsymbol{\zeta}} \log\left|\det J_{T^{-1}}(\boldsymbol{\zeta})\right|}_{1\times p}\right] \quad \text{(F.2.2a)}$$

$$\nabla_{\widetilde{\boldsymbol{\omega}}}\mathcal{L} = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\boldsymbol{I})}\left[\left(\underbrace{\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\theta})}_{1\times p}\underbrace{\nabla_{\boldsymbol{\zeta}} T^{-1}(\boldsymbol{\zeta})}_{p\times p} + \underbrace{\nabla_{\boldsymbol{\zeta}} \log\left|\det J_{T^{-1}}(\boldsymbol{\zeta})\right|}_{1\times p}\right)\underbrace{\nabla_{\widetilde{\boldsymbol{\omega}}} S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})}_{p\times p}\right] + \mathbf{1} \quad \text{(F.2.2b)}$$

where we have defined $\widetilde{\boldsymbol{\omega}} = (\widetilde{\omega}_1, \ldots \widetilde{\omega}_p) \in \mathbb{R}^p$ as the element-wise log of the variational standard deviations, $\widetilde{\omega}_p = \log \widetilde{\sigma}_p$. This transformation gives $\widetilde{\boldsymbol{\omega}}$ unbounded real-valued support.

In the case of categorical GLMS (Eq. (1.1.1)), the model parameter is given by $\boldsymbol{\theta} = \text{vec}(\boldsymbol{B})$. Here the model parameter $\boldsymbol{\theta}$ already has unconstrained real-valued support, so the ADVI gradients (F.2.2) simplify greatly. Since T is the identity function, we have $J_{T^{-1}}(\boldsymbol{\zeta}) = \nabla_{\boldsymbol{\zeta}} T^{-1}(\boldsymbol{\zeta}) = \boldsymbol{I}_p$ and $\nabla_{\boldsymbol{\zeta}} \log\left|\det J_{T^{-1}}(\boldsymbol{\zeta})\right| = \mathbf{0}_p$. Therefore, the ADVI gradients become

$$\nabla_{\widetilde{\boldsymbol{\mu}}}\mathcal{L} = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\boldsymbol{I})}\left[\underbrace{\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\theta})}_{1\times p}\right] \quad \text{(F.2.3a)}$$

$$\nabla_{\widetilde{\boldsymbol{\omega}}}\mathcal{L} = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};\mathbf{0},\boldsymbol{I})}\left[\underbrace{\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\theta})}_{1\times p}\underbrace{\nabla_{\widetilde{\boldsymbol{\omega}}} S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})}_{p\times p}\right] + \mathbf{1}_p \quad \text{(F.2.3b)}$$

If we take S_λ to be an elliptical standardization (Kucukelbir et al., 2017), we obtain

$$\nabla_{\widetilde{\boldsymbol{\omega}}} S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon}) = \text{diag}(\boldsymbol{\epsilon}^T \exp(\widetilde{\boldsymbol{\omega}})) = \begin{bmatrix} \epsilon_1 \exp(\widetilde{\omega}_1) & & \\ & \ddots & \\ & & \epsilon_p \exp(\widetilde{\omega}_p) \end{bmatrix}$$

So (stochastic) gradient steps to optimize the ADVI ELBO are conceptually straightforward to compute using Monte Carlo sampling and automatic differentiation of the joint density with respect to the parameter $\boldsymbol{\theta}$. However, the generic framework comes at a price. Whereas CAVI provides exact analytic solutions to each coordinate ascent update, ADVI must chase gradients, and these gradients are stochastic. As we will see, this can slow down inference; moreover, ADVI introduces multiple optimization hyperparameters (learning rate, number of Monte Carlo samples, and more). For a given problem, finding appropriate values of these hyperparameters can be challenging.

F.3 Gibbs sampling for multi-logit regression with Pòlya-Gamma augmentation

F.3.1 Bayesian Binomial Regression with Pòlya-Gamma augmentation

Here we derive the complete conditionals for Bayesian Binomial Regression with Pòlya-Gamma augmentation . Bayesian logistic regression is a special case, and Bayesian multiclass logistic regression is an extension (see Sec. F.3.2). This derivation will be useful for Sec. F.4, where we demonstrate the lack of closed-form CAVI for softmax regression with Pòlya-Gamma augmentation .

Our derivation largely follows the simple derivation given in Section 2 of (Choi et al., 2013), but provides some extra detail.[5] For the derivation, recall the density of a PG(b, c)

[5] We also make the generalization from logistic to binomial regression explicit. Although the tweak is straightforward, this expression is nicely more general and is also something we use when we handle the stick-breaking multi-class logistic regression.

random variable (Polson et al., 2013):

$$f(x; b, c) = \cosh^b(c/2) e^{-\frac{c^2}{2}x} g(x; b, 0) \tag{F.3.1}$$

where $h(\omega) := g(x; b, 0)$ is the density of a PG$(b, 0)$ random variable

$$g(x; b, 0) = \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)}{\Gamma(n+1)} \frac{(2n+b)}{\sqrt{2\pi x^3}} e^{-\frac{(2n+b)^2}{8x}} \tag{F.3.2}$$

So f is constructed from g via an *exponential tilt* and a renormalization. Note that to derive the complete conditionals, we will not need the form of $g(x; b, 0)$ anywhere in the derivation; we merely use (F.3.1).

Proposition F.3.1. *For the Bayesian Binomial Regression Model*[6] *with Pòlya-Gamma augmentation*

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ y_i \mid \boldsymbol{\beta} &\overset{ind}{\sim} Binomial\left(n_i, \frac{\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_i^T\boldsymbol{\beta}\}}\right), \quad i = 1, ..., N \\ \omega_i \mid \boldsymbol{\beta} &\overset{ind}{\sim} PG(n_i, \boldsymbol{x}_i^T\boldsymbol{\beta}), \quad i = 1, ..., N \end{aligned} \tag{F.3.3}$$

the complete conditional distributions are

$$(\omega_i \mid \boldsymbol{\beta}) \sim PG(n_i, \boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{F.3.4}$$

$$(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\omega}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}) \tag{F.3.5}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \left(\boldsymbol{X}^T\boldsymbol{\Omega}_{\boldsymbol{\omega}}\boldsymbol{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1} \tag{F.3.6}$$

$$\boldsymbol{\mu}_{\boldsymbol{\omega}} = \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\left(\boldsymbol{X}^T\boldsymbol{\kappa} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right) \tag{F.3.7}$$

where

$$\begin{aligned} \boldsymbol{\kappa} &= (y_1 - n_1/2, ..., y_N - n_N/2) \\ \boldsymbol{\Omega}_{\boldsymbol{\omega}} &\text{ is the diagonal matrix of } \boldsymbol{\omega}_i\text{'s} \end{aligned} \tag{F.3.8}$$

Proof. That (F.3.4) is the complete conditional for ω_i follows immediately from the conditional independence of $\boldsymbol{\omega}$ and $\boldsymbol{y}$ in the model. In particular, the posterior density is given by

$$p(\boldsymbol{\beta}, \boldsymbol{\omega} \mid \boldsymbol{y}) \propto \left[\prod_{i=1}^{N} p(y_i \mid \boldsymbol{\beta}) p(\omega_i \mid \boldsymbol{\beta})\right] p(\boldsymbol{\beta}) \tag{F.3.9}$$

Hence, clearly,

$$p(\omega_i \mid \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{\omega}_{-i}) \propto p(\omega_i \mid \boldsymbol{\beta})$$

[6]Note that we use N to denote the number of observations and n_i to denote the number of binomial trials per observation.

It remains to show that (F.3.5) is the complete conditional for $\boldsymbol{\beta}$

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \boldsymbol{y}) &\propto \left[\prod_{i=1}^{N} p(y_i \mid \boldsymbol{\beta}) p(\omega_i \mid \boldsymbol{\beta})\right] p(\boldsymbol{\beta}) \\ &\overset{1}{\propto} \left[\prod_{i=1}^{N} \frac{e^{(\boldsymbol{x}_i^T\boldsymbol{\beta})y_i}}{\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{n_i}}\right] \left[\cosh^{n_i}\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}}{2}\right) e^{-\frac{1}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\omega_i} h(\omega_i)\right] p(\boldsymbol{\beta}) \\ &\overset{2}{\propto} \left[\prod_{i=1}^{N} \frac{e^{(\boldsymbol{x}_i^T\boldsymbol{\beta})y_i}}{\cancel{\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{n_i}}}\right] \left[\frac{\cancel{\left(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)^{n_i}}}{2^{n_i} e^{\frac{1}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta})n_i}} e^{-\frac{1}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\omega_i} \cancel{h(\omega_i)}\right] p(\boldsymbol{\beta}) \\ &\overset{3}{\propto} p(\boldsymbol{\beta}) \exp\left\{\sum_{i=1}^{N} (y_i - \frac{n_i}{2})(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \frac{\omega_i}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right\} \end{aligned} \tag{F.3.10}$$

where (1) fills in forms for densities (using $h(\omega) := g(\omega; 1, 0)$), (2) uses that $\cosh(z) = \frac{1+e^{2z}}{2e^z}$ (and absorbs $h(\omega_i)$ and 2^{-n_i} into the constant of proportionality), and (3) reveals that the complete conditional is Gaussian.

To obtain an explicit form for the multivariate Gaussian, we need what (Choi et al., 2013) calls a routine Bayesian regression-type calculation:

$$p(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \boldsymbol{y}) \propto p(\boldsymbol{\beta}) \exp\left\{\sum_{i=1}^{N} (y_i - \frac{n_i}{2})(\boldsymbol{x}_i^T\boldsymbol{\beta}) - \frac{\omega_i}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right\}$$

setting $\kappa_i := y_i - \frac{n_i}{2}$

$$\overset{1}{\propto} p(\boldsymbol{\beta}) \exp\left\{\sum_{i=1}^{N} -\frac{\omega_i}{2}\left(\boldsymbol{x}_i^T\boldsymbol{\beta} - \frac{\kappa_i}{\omega_i}\right)^2\right\}$$

defining $\boldsymbol{z} := (\frac{\kappa_1}{\omega_1}, ..., \frac{\kappa_N}{\omega_N})$ and $\boldsymbol{\Omega} := \text{diag}(\omega_1, ..., \omega_N)$

$$\begin{aligned} &\overset{2}{\propto} p(\boldsymbol{\beta}) \exp\left\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})\right\} \\ &\overset{3}{\propto} p(\boldsymbol{\beta}) \exp\left\{-\frac{1}{2}(\boldsymbol{X}^+\boldsymbol{z} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}^+\boldsymbol{z} - \boldsymbol{\beta})\right\} \end{aligned} \tag{F.3.11}$$

where (1) is by completing the square, (2) writes the weighted sum of squares in matrix notation, and (3) isolates $\boldsymbol{\beta}$, using $\boldsymbol{X}^+$, the Moore-Penrose psuedo-inverse of $\boldsymbol{X}$.[7]

Thus, we see that $p(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \boldsymbol{y})$ is proportional to the product of two multivariate Gaussians: $p(\boldsymbol{\beta})$, which has mean $\boldsymbol{\mu}_0$ and covariance $\boldsymbol{\Sigma}_0$, and another Gaussian, which has mean $\boldsymbol{X}^+\boldsymbol{z}$ and covariance $(\boldsymbol{X}^T\boldsymbol{\Omega}\boldsymbol{X})^{-1}$. We know from the exponential family representation of the Gaussian that the result can be obtained by summing at the scale of natural parameters – which for the Gaussian are the precision and precision-weighted mean. Using this, we obtain

$$p(\boldsymbol{\beta} \mid \boldsymbol{\omega}, \boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$$

[7]Specifically, since $\boldsymbol{X}\boldsymbol{X}^+ = \boldsymbol{I}$, we use

$$\begin{aligned} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Omega}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z})^T\boldsymbol{\Omega}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{z}) \\ &= \left(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{z})\right)^T\boldsymbol{\Omega}\left(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{z})\right) \\ &= (\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{z})^T\boldsymbol{X}^T\Omega\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{X}^+\boldsymbol{z}) \end{aligned}$$

.

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\omega}} &= \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{\Omega}_{\boldsymbol{\omega}}\boldsymbol{X}\right)^{-1} \\ \boldsymbol{\mu}_{\boldsymbol{\omega}} &= \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{\Omega}_{\boldsymbol{\omega}}\boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{z}\right) \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\omega}}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{\kappa}\right)\end{aligned}$$

□

Remark F.3.1. In the special case where $n_i \equiv 1$, Bayesian binomial regression reduces to Bayesian logistic regression. △

F.3.2 Bayesian multi-logit regression with Pòlya-Gamma augmentation

(Held & Holmes, 2006) show that for multi-class logistic regression with the standard, canonical (multi-logit) link, the conditional likelihood $L(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \boldsymbol{\beta}_{-k})$ over categorical outcomes $\boldsymbol{y} \in \{1, \ldots, K\}^N$ has the form of a logistic regression on the class indicators $\widehat{y}_{ik} \in \{0, 1\}$. This observation motivates the conversion of Bayesian multinomial regression into a conditionally conjugate model. We present the complete conditionals here as they are used to construct a Gibbs sampler in the experiments (see Sec. G.3.2). However, in Sec. F.4 we show that the construction does not yield closed-form CAVI updates (as reported in row 3 of Table 1).

First, following (Held & Holmes, 2006), note that we can represent the complete conditionals for $\boldsymbol{\beta}_k, k = 1, ..., K-1$ in terms of the conditional likelihoods $L(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \boldsymbol{\beta}_{-k})$

$$p(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \boldsymbol{\beta}_{-k}) \propto p(\boldsymbol{\beta}_k)\, L(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \boldsymbol{\beta}{-}k)$$

where the conditional likelihoods satisfy

$$\begin{aligned}L(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \boldsymbol{\beta}_{-k}) &\propto \prod_{i=1}^{N}\prod_{k=1}^{K} p_{ik}^{\widehat{y}_{ik}} \\ &\propto \prod_{i=1}^{N} (\gamma_{ik})^{\widehat{y}_{ik}}(1-\gamma_{ik})^{1-\widehat{y}_{ik}}\end{aligned}$$

where

$$\begin{aligned}\gamma_{ik} &= \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik})}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik})} \qquad \text{(F.3.12)} \\ C_{ik} &:= \log\sum_{j\neq k}\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_j)\end{aligned}$$

which reveals that the conditional likelihood $L(\boldsymbol{\beta}_k \mid \boldsymbol{y}, \boldsymbol{\beta}_{-k})$ has the form of a logistic regression on class indicators $\widehat{y}_{ik}$.

The form of (F.3.12) and the success of Pòlya-Gamma augmentation with standard (binary) logistic regression suggests that we should construct an augmented model for Bayesian multi-class logistic regression by taking, for $i = 1, ..., N$ and $k = 1, ...K-1$,

$$\omega_{ik} \mid \boldsymbol{\beta}_k \overset{\text{ind}}{\sim} \text{PG}(1, \boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik}),$$

a slight tweak on the construction for standard (binary) logistic regression (F.3.3), where we had $\omega_i \mid \boldsymbol{\beta} \overset{\text{ind}}{\sim} \text{PG}(1, \boldsymbol{x}_i^T\boldsymbol{\beta})$.

Following (F.3.10), but using the conditional likelihood and altered construction for the Pòlya-Gamma auxiliary variables[8] we find

$$\begin{aligned}&p(\boldsymbol{\beta}_k \mid \boldsymbol{\omega}, \boldsymbol{y}, \boldsymbol{\beta}_{-k}) \\ &\propto \left[\prod_{i=1}^{N}\frac{e^{(\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik})\widehat{y}_{ik}}}{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik}})}\right] \\ &\qquad\times\left[\cosh\left(\frac{\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik}}{2}\right)e^{-\frac{1}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik})^2\omega_{ik}}h(\omega_{ik})\right]p(\boldsymbol{\beta}_k) \\ &\propto \left[\prod_{i=1}^{N}\frac{e^{(\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik})\widehat{y}_{ik}}}{\cancel{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik}})}}\right] \\ &\qquad\times\left[\frac{\cancel{(1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik}})}}{\cancel{e^{\frac{1}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik})}}}e^{-\frac{1}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta}-C_{ik})^2\omega_{ik}}\cancel{h(\omega_{ik})}\right]p(\boldsymbol{\beta}_k) \\ &\propto p(\boldsymbol{\beta}_k)\exp\left\{\sum_{i=1}^{N}(\widehat{y}_{ik}-\frac{1}{2})(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik}) - \frac{\omega_{ik}}{2}(\boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik})^2\right\}\end{aligned} \qquad \text{(F.3.13)}$$

Continuing to parallel the argument of Section F.3.1, using Eq. (F.3.13) instead of Eq. (F.3.10), we find that the complete conditionals are given by

$$\boldsymbol{\beta}_k \mid \boldsymbol{\omega}_k, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad \text{(F.3.14a)}$$

$$\omega_{ik} \mid \boldsymbol{\beta}_k \sim \text{PG}(1, \boldsymbol{x}_i^T\boldsymbol{\beta}_k - C_{ik}) \qquad \text{(F.3.14b)}$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{X}^T\boldsymbol{\Omega}_k\boldsymbol{X}\right)^{-1} \\ \boldsymbol{\mu}_{\boldsymbol{\omega}} &= \boldsymbol{\Sigma}_k\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{X}^T\boldsymbol{\Omega}_k\boldsymbol{z}_k\right)\end{aligned}$$

for

$$\begin{aligned}\boldsymbol{\Omega}_k &= \text{diag}\left(\omega_{1k}, \ldots, \omega_{Nk}\right) \\ \boldsymbol{z}_k &= \begin{bmatrix}\frac{\widehat{y}_{1k}-1/2}{\omega_{1k}} + C_{1k} \\ \vdots \\ \frac{\widehat{y}_{Nk}-1/2}{\omega_{Nk}} + C_{Nk}\end{bmatrix}\end{aligned}$$

A valid Gibbs sampler is obtained by iteratively sampling from Eqs. (F.3.14a) and (F.3.14b).

F.4 Lack of closed-form CAVI for Bayesian multi-logit regression and Pòlya-Gamma augmentation

Here we demonstrate the lack of closed-form CAVI for Bayesian multi-logit regression under Pòlya-Gamma augmentation (as reported in row 3 of Table 1). To do so, we focus on the complete conditionals for ω_{ik}. Namely, if we would like to perform coordinate ascent

[8] Note that, for this example, we are unnecessarily restricting to the case of multiclass logistic regression ($n_i \equiv 1$). The same argument that we make here would of course also hold for multinomial regression, which is a generalization.

variational inference (CAVI), we parallel the argument of (E.3.8) , but with our current multi-class situation whereby we work with the complete conditional for ω_{ik} as $\text{PG}(1, c_{ik})$, where $c_{ik} = \boldsymbol{x}_i^T \boldsymbol{\beta}_k - C_{ik}$. This differs slightly from the standard (binary) logistic regression case, where the complete conditional for ω_i was $\text{PG}(1, c_i)$, where $c_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. In the current case, we find by paralleling (E.3.8) that we eventually need

$$\begin{aligned}\mathbb{E}_{q_{-\boldsymbol{\omega}}}[c_{ik}]^2 &= \left(\boldsymbol{x}_i^T \mathbb{E}_{q_{\boldsymbol{\beta}_k}}[\boldsymbol{\beta}_k] - \mathbb{E}_{q_{-\omega_{ik}}}[C_{ik}]\right)^2 \\ &= \left(\boldsymbol{x}_i^T \mathbb{E}_{q_{\boldsymbol{\beta}_k}}[\boldsymbol{\beta}_k] - \mathbb{E}_{q_{\boldsymbol{\beta}_{-k}}}[\log \sum_{j \neq k} \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}_j)]\right)^2\end{aligned} \tag{F.4.1}$$

and while the first expectation in equation (F.4.1) is straightforward and parallels what we computed in binary logistic regression, the expected log-sum-exp is distinct to the multiclass case, and has no closed form. Indeed, the expected log-sum-exp is a notorious blocker to closed-form CAVI. Indeed, precisely this fact motivated Delta Variational Inference (Braun & McAuliffe, 2010) (Wang & Blei, 2013). For an enumeration of many bounds to this expression, see (Depraetere & Vandebroek, 2017).

Thus, if one seeks variational inference with closed-form updates, Pòlya-Gamma augmentation solves the problem for Bayesian binomial regression, but not for Bayesian multi-class logistic regression (or more generally Bayesian multinomial regression), at least not when using the standard canonical (multi-logit) link.

F.5 Stick-breaking multi-class logistic regression

The stick-breaking construction of the multi-class logistic regression regression (Linderman et al., 2015) is useful for the purpose of exploiting Pòlya-Gamma augmentation for efficient inference. First, the density of a *categorical* distribution over K categories with parameter $\boldsymbol{\pi} = (\pi_1, ..., \pi_k) \in \Delta_{K-1}$ can be represented in a stick-breaking manner as a product of $K-1$ Bernoullis. The density can be expressed as

$$\prod_{k=1}^{K-1} \widetilde{\pi_k}^{\widehat{y}_{ik}} (1 - \widetilde{\pi_k})^{1 - \widehat{y}_{ik}} \tag{F.5.1}$$

where $\widetilde{\pi_k} := \dfrac{\pi_k}{1 - \sum_{j<k} \pi_j}$ is the Bernoulli parameter and $\widehat{y}_{ik} = 1$ if the ith observation is the kth category, and $\widehat{y}_{ik} = 0$ otherwise.

Stick-breaking multi-class logistic regression uses Eq. F.5.1 to construct a multi-class logistic regression over K categories as a product of $K-1$ logistic regressions

$$\prod_{k=1}^{K-1} \left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}}\right)^{\widehat{y}_{ik}} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}}\right)^{1 - \widehat{y}_{ik}}$$

The multinomial parameter $\boldsymbol{\pi}_i$ has explicit form given by

$$\pi_{ik} = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}} \prod_{j<k} \frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_j^{\text{SB}}}}, \quad k = 1, ..., K \tag{F.5.2}$$

where $\boldsymbol{\beta}_K \equiv 0$.

Label asymmetry. The stick-breaking formulation induces a label asymmetry, which can complicate prior-setting and reduce representational capacity (Zhang & Zhou, 2017). For example, consider the case of multiclass logistic regression (so $n_i = 1$). In standard multinomial regression, we have

$$p(\widehat{y}_{ik} = 1 \mid \boldsymbol{\beta}^{\text{ML}}) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{ML}}}}{1 + \sum_{k=1}^{K-1} e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{ML}}}} \tag{F.5.3}$$

whereas in stick-breaking multinomial regression, we have (via Eq. F.5.2)

$$P(\widehat{y}_{ik} = 1 \mid \boldsymbol{\beta}^{\text{SB}}) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_k^{\text{SB}}}} \prod_{j<k} \frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}_j^{\text{SB}}}}$$

which differs from (F.5.3) in that it clearly imposes fewer geometric constraints on the classification decision boundaries for smaller k. For instance, p_{i1} can be larger than 50 % if $\boldsymbol{x}_i^T \boldsymbol{\beta}_1 > 0$, whereas p_{i2} can be larger than 50 % only if $\boldsymbol{x}_i^T \boldsymbol{\beta}_1 < 0$ and $\boldsymbol{x}_i^T \boldsymbol{\beta}_2 > 0$. Because of label asymmetry, predictive performance can be sensitive to how the K different categories are ordered. The geometric constraints implied by any given ordering of the labels can cause the model to struggle to learn the true decision boundaries.

G Supplemental information for experiments

Open-source python code for reproducing experiments can be found at https://github.com/tufts-ml/categorical-from-binary.

G.1 Data simulations

G.1.1 Dataset generation

We generate simulated datasets from a categorial distribution with a softmax (multi-logit) inverse link function and given specifications (N samples, K categories, M covariates). For a given context (N, K, M), we may generate D different datasets by setting the random seed to a different value.

First, we generate covariate matrices $\boldsymbol{X} \in \mathbb{R}^{N \times M}$ such that all entries are drawn i.i.d from $\mathcal{N}(0, 1)$. We use $\boldsymbol{x}_i$ to refer to the ith row of $\boldsymbol{X}$ for $i = 1, \ldots, N$.

Next, we draw regression weights $\boldsymbol{B} \in \mathbb{R}^{(M+1) \times K}$ in a way that allows us to control category predictability. We describe how to sample entries β_{mk} for $m = 0, 1, \ldots, M$ and $k = 1, \ldots, K$. We begin by generating intercepts for

each category by sampling $\beta_{0k} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\text{int}})$. For covariates $m = 1, \ldots, M$, we draw $\beta_{mk} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2_{mk})$, where

$$\sigma^2_{mk} = \begin{cases} \sigma^2_{\text{high}}, & \text{if } k = \lceil m/S \rceil \\ \sigma^2_{\text{low}}, & \text{otherwise} \end{cases}$$

for $\sigma^2_{\text{high}} > \sigma^2_{\text{low}}$ and $S := \lfloor M/K \rfloor$. The motivation is as follows: We partition the M covariates into $K+1$ covariate groups. Covariate groups $k = 1, ..., K$ each have S members that are potentially predictive of the kth category. Such covariates have regression entries $\beta_{mk} \sim \mathcal{N}(0, \sigma^2_{\text{high}})$; the relatively high variance $\sigma^2_{\text{high}} > \sigma^2_{\text{low}}$ allows the regression coefficients to escape the mean of zero. As the value of σ^2_{high} increases, the overall predictability of the categories given the covariates increases. Note that there may be an additional $(k = 0)$th group, with $M \bmod K$ members, which is not predictive of a specific category.

Finally, we generate categorical observations by associating the covariates and regression weights via the softmax (multi-logit) inverse link function.

Overall, our data generating mechanism is

$$\begin{aligned} x_{im} &\overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \qquad i = 1, \ldots, N, \; m = 1, \ldots, M \\ \beta_{mk} &\overset{\text{ind}}{\sim} N(0, \sigma^2_{mk}), \quad m = 0, \ldots, M, k = 1, \ldots, K \\ &\text{where } \sigma^2_{mk} = \begin{cases} \sigma^2_{\text{int}}, & \text{if } m = 0 \\ \sigma^2_{\text{high}}, & \text{if } m \geq 1 \text{ and } k = \lceil m/S \rceil \\ \sigma^2_{\text{low}}, & \text{otherwise} \end{cases} \\ y_i \mid \boldsymbol{x}_i, \boldsymbol{B} &\sim \text{Softmax}(\boldsymbol{B}^T \dot{\boldsymbol{x}}_i) \end{aligned}$$

for observations $i = 1, \ldots, N$, covariates $m = 1, \ldots, M$ and categories $k = 1, \ldots, K$, and where $\dot{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^T)^T$ are the covariates prepended with a value of 1 to correspond to the intercept term.

Unless otherwise specified, we fix $\sigma^2_{\text{low}} = 0.001$ and $\sigma^2_{\text{int}} = 0.25$. We vary σ^2_{high} throughout the experiments to control predictability.

G.1.2 Metrics

To estimate the predictability of categories, we estimate the mean covariate-conditional category entropy for each dataset:

$$\mathbb{E}_X \mathbb{H}[Y \mid X] \approx -\sum_{i=1}^{N} \sum_{k=1}^{K} p(y_i = k \mid \boldsymbol{x}_i, \boldsymbol{B}) \log p(y_i = k \mid \boldsymbol{x}_i, \boldsymbol{B}) \tag{G.1.1}$$

where p refers to the category probabilities and $\boldsymbol{B}$ refers to the known regression weights from the true data generating process (softmax).

The mean holdout log-likelihood for the rth prediction method is computed by:

$$\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \log p_r(y_i = k \mid \boldsymbol{x}_i, \boldsymbol{B}_r) \tag{G.1.2}$$

where the category probability formula p_r and point estimate for $\boldsymbol{B}_r$ are determined by the values of the corresponding columns for the rth row of Table G.2.[9]

G.2 Bayesian model averaging experiment: Supplemental information

G.2.1 Methodology

Data generation. We generated multiple datasets from a categorial distribution with a softmax (multi-logit) using the technique described in Sec. G.1.1. In particular, we randomly generated 16 datasets by taking the number of categories to be $K \in \{3, 10\}$, the number of covariates to be a multiple of the number of categories via $M = aK$ for $a \in \{1, 2\}$, the number of samples to be a multiplier on the number of parameters via $N = bP$ for $b \in \{10, 20, 40, 80, 160\}$ (where recall that the number of parameters is given by $P = K(M + 1)$ due the presence of an intercept), and $\sigma^2_{\text{high}} \in \{0.1, 4.0\}$ to control the predictability of the categorical observations.

Training. For each dataset, we used 80% of the data for model training and held out the remaining 20% for evaluation. We applied our IB-CAVI inference technique with the logit link (so H was taken as the standard logistic cdf). For each dataset, we ran IB-CAVI until the surrogate lower bound $\texttt{ELBO}_{\text{IB}}$ had a mean value (across samples and categories) that dropped by 0.1 or less on consecutive iterations.

Predictive likelihoods. Recall that we have partitioned each dataset into training data $\boldsymbol{y}^{\text{train}} \in \{1, ..., K\}^{N_{\text{train}}}$ and hold-out test data $\boldsymbol{y}^{\text{test}} \in \{1, ..., K\}^{N_{\text{test}}}$. After training the model on $\boldsymbol{y}^{\text{train}}$, we consider three different predictive likelihoods for test set observations y_i^{test}, $i \in 1, ..., N_{\text{test}}$. In particular, we can compute $p_{\text{CBC}}(y_i^{\text{test}} \mid \boldsymbol{y}_{\text{train}})$, $p_{\text{CBM}}(y_i^{\text{test}} \mid \boldsymbol{y}_{\text{train}})$, and $p_{\text{BMA}}(y_i^{\text{test}} \mid \boldsymbol{y}_{\text{train}})$. The former two quantities are estimated by substituting IB-CAVI's variational posterior expectation into the relevant model's category probability formulae, Eqs. 2.4.1 and 2.4.2. The latter quantity is computed from the former two quantities via Eq. (4.0.1) using the method of Sec. 4.

[9] Two of the modeling strategies - namely `Softmax (via MLE)` and `Baserate frequency` - can produce predictive probabilities of exact or numerical zero, e.g. when a category is observed in the test set that was never observed in the training set. A single such instance will drive the log-likelihood metric to $-\infty$ regardless of the log likelihoods for any other sample. To handle this issue, we renormalize these models to produce a minimum predictive probability of $\epsilon := 10^{-10}$ for each category.

Discrepancy from true category probabilities. Since the data is simulated, we have access to the "ground truth" predictive likelihood $p_{\text{true}}(y_i^{\text{test}} \mid \boldsymbol{y}_{\text{train}})$ for each test set sample, obtained by substituting the true regression weights $\boldsymbol{B}_{\text{true}}$ into the softmax likelihood. We can therefore evaluate the performance of our three estimated predictive likelihoods by computing the discrepancy between each approximation and this ground truth:

$$d_i := D_{\text{KL}}\left[p_{\text{true}}(y_i^{\text{test}} \mid \boldsymbol{y}_{\text{train}}) \parallel p_{\mathcal{M}}(y_i^{\text{test}} \mid \boldsymbol{y}_{\text{train}})\right]$$

where $\mathcal{M} \in \{\text{CBC}, \text{CBM}, \text{BMA}\}$, D_{KL} is the Kullback-Leibler divergence, and $i = 1, ..., N_{\text{test}}$. Our performance measure for each estimated predictive likelihood is then the mean discrepancy across the test set, i.e. $\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} d_i$.

G.2.2 Results

.

Figure G.1 provides an expanded version of Figure 1. Some patterns of interest:

- As the number of categories and covariates and predictability $(K, M, \sigma^2_{\text{high}})$ are fixed, the error in the IB approximation decreases as the number of samples N increases (as a multiple $b \in \{10, 20, 40, 80, 120\}$ on the number of parameters).
- As the predictability of the categorical response (σ^2_{high}) increases, the CBC model becomes better than CBM at serving as a target of the approximation. (To see this, compare the left column to the right column in Fig. 1.) Since the predictability of the dataset may not be known in advance, this fact might seem to create a difficult model selection problem. Luckily, the Bayesian model averaging (BMA) tracks the relative appropriateness of each model change by toggling the weight on the CBC model w_{CBC}.
- The relative advantage of the CBC model over the CBM model also seems to increase as the number of parameters $P = K(M + 1)$ increases. (To see this, compare the top rows to the bottom rows in Fig. G.1.)

Table G.1 provides more detailed information about the results shown in Fig. G.1.

G.3 Variational Bayes vs. Maximum Likelihood: Supplemental information

G.3.1 Methodology

We generate data using the method described in Sec. G.1.1. For this experiment, we generate $D = 10$ datasets per simulation context, which is a particular choice of $K = 3, M = 2K, N \in \{1, 100\} * P$ where $P = K(M + 1)$, $\sigma_{\text{high}} \in \{0.01, 0.5, 1, 2, 5, 10, 20, 50, 100\}, \sigma_{\text{low}} = 0.01, \sigma_{\text{int}} = 1.0$. The choice of $M = 2K$ could be imagined as the number of covariates under a light (order 2) autoregressive structure. For each dataset, we used 80% of the data for model training and held out the remaining 20% for evaluation.

G.3.2 Modeling Strategies

We compare a number of different modeling strategies:

1. *Data generating process*: We take the known regression coefficients $\boldsymbol{B}_{\text{true}}$ and plug it into the softmax (multi-logit) categorical probability function.
2. *Softmax (via MLE)*: We estimate the MLE, $\boldsymbol{B}_{\text{MLE}}$, for and softmax a.k.a. multi-logit model. The optimization was computed using automatic differentiation in `jax` (and default convergence parameters). We can interpret the results of the optimization as an (approximate) MLE due to the convexity of the multi-logit function. The solver used was BFGS, which is the only solver that `jax` supports[10]. We can make predictions on new samples by plugging in $\boldsymbol{B}_{\text{MLE}}$ to the multi-logit categorical probability function.
3. CB *(via IB-CAVI)*: We compute CAVI for CB models (specifically, the CBC-Probit and CBM-Probit) with a $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ prior using the variational technique with independent binary approximation described in the main body of the text. We concluded convergence when the drop in the mean ELBO (with the mean taken across the number of samples N and categories K) was less than 0.1 across consecutive iterations. The variational posterior mean $\mathbb{E}_q[\boldsymbol{B}]$ was used as a point estimate for $\boldsymbol{B}$, and then substituted into the category probability formula for either the CBC-Probit or CBM-Probit.
4. *Baserate frequency*: We use the raw frequencies of each category in the training set and use those as the predicted category probabilities for test set data, regardless of the value of the covariates, i.e.
$$p(y_i = k \mid \boldsymbol{x}_i) = f_k \tag{G.3.1}$$
where f_k is the frequency with which the kth category was observed in the training set.

The differences between the modeling strategies are summarized in Table G.2.

Training. The MLE and IB-CAVI were both initialized at the zero matrix. For each dataset, we ran IB-CAVI until the surrogate lower bound $\texttt{ELBO}_{\text{IB}}$ had a mean value (across samples and categories) that dropped by 0.1 or less on consecutive iterations.

[10] as of documentation revision 1182e7aa

Table G.1: *Additional results from applying the approximate Bayesian Model Averaging technique of Sec. 4 to simulated datasets.* This table provides more detailed information about the analysis depicted in Fig. 1. w_{CBC} gives the weight that the technique assigns to the CBC model.

N	K	M	σ_{high}	w_{CBC}	Mean KL divergence to true probabilities from: CBM	CBC	BMA
120	3	3	0.100	0.021	0.016	0.017	0.015
120	3	3	2.000	0.861	0.067	0.046	0.040
240	3	3	0.100	0.005	0.031	0.065	0.031
240	3	3	2.000	0.999	0.043	0.027	0.026
480	3	3	0.100	0.002	0.007	0.021	0.007
480	3	3	2.000	1.000	0.023	0.018	0.018
960	3	3	0.100	0.000	0.002	0.018	0.002
960	3	3	2.000	1.000	0.029	0.020	0.020
1920	3	3	0.100	0.000	0.002	0.008	0.002
1920	3	3	2.000	1.000	0.030	0.018	0.018
210	3	6	0.100	0.001	0.030	0.069	0.030
210	3	6	2.000	1.000	0.116	0.093	0.093
420	3	6	0.100	0.001	0.016	0.030	0.016
420	3	6	2.000	1.000	0.090	0.030	0.030
840	3	6	0.100	0.000	0.009	0.033	0.009
840	3	6	2.000	1.000	0.071	0.023	0.023
1680	3	6	0.100	0.000	0.007	0.025	0.007
1680	3	6	2.000	1.000	0.077	0.018	0.018
3360	3	6	0.100	0.000	0.001	0.017	0.001
3360	3	6	2.000	1.000	0.076	0.014	0.014
1100	10	10	0.100	0.002	0.042	0.054	0.042
1100	10	10	2.000	1.000	0.114	0.057	0.057
2200	10	10	0.100	0.008	0.028	0.034	0.028
2200	10	10	2.000	1.000	0.098	0.052	0.052
4400	10	10	0.100	0.023	0.011	0.013	0.011
4400	10	10	2.000	1.000	0.079	0.017	0.017
8800	10	10	0.100	0.925	0.008	0.008	0.008
8800	10	10	2.000	1.000	0.074	0.017	0.017
17600	10	10	0.100	1.000	0.004	0.004	0.004
17600	10	10	2.000	1.000	0.073	0.017	0.017
2100	10	20	0.100	0.005	0.041	0.051	0.041
2100	10	20	2.000	1.000	0.135	0.061	0.061
4200	10	20	0.100	0.000	0.021	0.026	0.021
4200	10	20	2.000	1.000	0.128	0.053	0.053
8400	10	20	0.100	0.006	0.012	0.015	0.012
8400	10	20	2.000	1.000	0.111	0.028	0.028
16800	10	20	0.100	1.000	0.006	0.007	0.007
16800	10	20	2.000	1.000	0.109	0.023	0.023
33600	10	20	0.100	1.000	0.004	0.004	0.004
33600	10	20	2.000	1.000	0.103	0.022	0.022

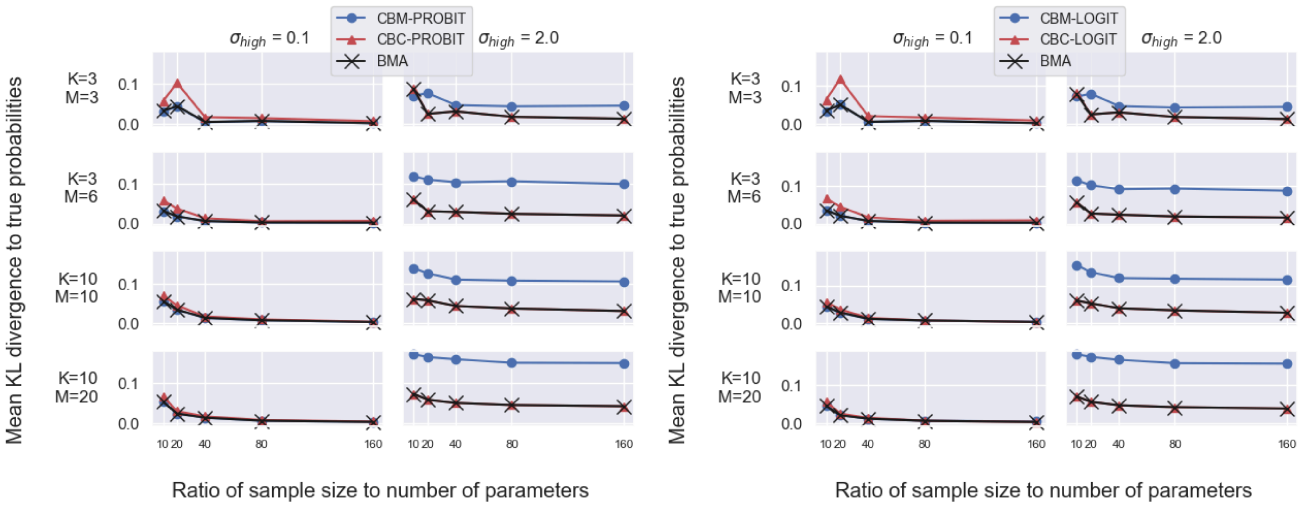

Figure G.1: **Determining a good target for the** IB **approximation.** Each point corresponds to a simulated dataset with some number of categories K and covariates M. Plotted are the mean KL divergences on hold-out test data to the true categorical probabilities for predictions of the CBMand CBCmodels (both estimated with IB-CAVI), as well as a Bayesian model averaging (BMA) of the two IB approximation targets. The left subplot use CB Probit models, whereas the right plot uses CB Logit models. Within each subplot, the level of predictability of the categorical responses is *weak* for datasets in the left column ($\sigma_{\text{high}} = 0.1$) and *strong* for those in the right column ($\sigma_{\text{high}} = 2.0$).

Table G.2: Summary of various modeling strategies for categorical data as used in the simulations experiment.

Modeling Strategy	**Model**	**Inference for B**
Data generating process	Softmax	B is known
Softmax (via MLE)	Softmax	MLE on softmax model
CBC-Probit (via IB-CAVI)	CBC	Variational posterior mean for the IB model, $\mathbb{E}_q[B]$
CBM-Probit (via IB-CAVI)	CBM	Variational posterior mean for the IB model, $\mathbb{E}_q[B]$
Baserate frequency	Equation (G.3.1)	N/A

Results. The confidence intervals in Fig. 2 were determined by bootstrapping.

G.4 Holdout performance over time: Supplemental information

The purpose of this experiment is to compare IB-CAVI against other methods for Bayesian inference with categorical GLMs. In particular, we compare the performance on holdout data as a function of training time.

G.4.1 Datasets

Simulated datasets. We construct simulated datasets using the method described in Sec. G.1.1 for given specifications (N samples, K categories, M covariates). We set $\sigma^2_{\text{high}} = 2.0$.

Real datasets. We investigate the following real datasets:

1. The *Detergent Purchase* dataset (Imai & Van Dyk, 2005), which has 2,657 observations, 6 covariates, and 6 categorical responses. Each record represents the purchase of a laundry detergent by a household in Sioux Falls, South Dakota. The prediction goal is to identify which of the 6 laundry detergents was purchased given the prices of all 6 detergents. This data is available via the GPL-3 license at `https://github.com/kosukeimai/MNP/blob/master/data/detergent.txt.gz`. For the original paper using this dataset, see (Chintagunta & Prasad, 1998).

2. The *Anuran Frog Calls* dataset (Colonna et al., 2017), which has 7,195 observations, 22 covariates, and 10 categorical responses. Each record represents extracted audio features (mel-frequency cepstrum coefficients (MFCCs) from a recording of a frog making some natural noises. The prediction goal is to identify which of 10 species the frog belongs to. This data is available from the UC-Irvine Machine Learning Repository via an open-access CC-BY license. For an in-depth paper using this data, see Colonna et al. (2016).

3. The *Glass Identification* dataset (German, 1987), which has 214 samples, 9 covariates, and 6 response categories that were observed. Each record represents observable properties of a physical sample of glass, with the prediction goal being to identify which of six types of glass the sample represents. A seventh possible category is noted in the data description but never observed. This data is available from the UC-Irvine Machine Learning Repository under an open-access CC-BY license. For the original paper using this dataset, see Evett & Spiehler (1987).

4. A *Single-User Process Start* dataset, which has 17,724 observations, 1,553 covariates, and 1,553 categorical responses. The dataset is constructed from the Comprehensive, Multi-Source Cybersecurity Events Dataset (Kent, 2015) using the methods of Sec. G.6. Each record contain one process start from one user account `U293@DOM1` along with the identity and timing of the $W = 5$ immediately preceding process starts. The prediction goal is to identify the next process start. The data is open-access with all copyrights waived, and the preprocessing used is available at `https://github.com/tufts-ml/categorical-from-binary`.

For all real datasets, we z-transformed all covariates, as the range of some variables is very small (e.g. consider the `RI` variable in the *glass identification* dataset, which only varies from 1.51 to 1.52). This lets us use independent $\mathcal{N}(0, 1)$ priors on the regression weights for each covariate-category combination. No missing data occurred in any of the datasets.

G.4.2 Modeling Strategies

Here we describe the various modeling strategies we used for Bayesian categorical regression modeling of the provided datasets. For motivation on which methods to include vs. exclude in the experiment, see the discussion of Sec. F.

1. CB-*Probit and* CB-*Logit (via IB-CAVI)*: We compute CAVI for CB-Probit and CB-Logit models with a $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ prior using the variational technique with independent binary approximation described in the main body of the text.

2. *Softmax regression (via automatic differentiation variational inference (ADVI)).* The gradient updates for softmax regression (whose parameters have support of unconstrained reals) are described in Sec. F.2. We implement these updates in `jax`, and optimize using Algorithm 1 of (Kucukelbir et al., 2017). We follow the recommendations of that paper to guide the optimization details: one Monte Carlo sample per update, and adaptive step-size sequences with varying learning rates but all other hyper-parameters kept at their recommended defaults.

3. *Softmax regression (via the No U-Turn Sampler (NUTS))* We sample from the posterior of softmax regression using the No U-Turn Sampler (NUTS) (Hoffman et al., 2014) as implemented in the Python package `numpyro`.

4. *Softmax regression (via Gibbs after Pòlya-Gamma augmentation)* Here we model the data

with softmax regression (more specifically the identified version of it which is obtained by setting $\boldsymbol{\beta}_K \equiv 0$; this is often called multi-logit regession), but using the Gibbs sampler which is available after Pòlya-Gamma augmentation . The complete conditionals for the Gibbs sampler are given in Sec. F.3.

G.4.3 General experimental methodology

For training, we used 80% of the data for model training and held out the remaining 20% for evaluation for all datasets except glass identification. Due to the small size of the glass identification dataset, we instead used a 90%/10% split. All methods were initialized to have their matrix of regression weights be $\boldsymbol{B} = \mathbf{0}$. Each inference method was run for a preset number of iterations (ADVI, IB-CAVI) or samples (NUTS, Gibbs) in an attempt to make the running time for each method similar. The performance of NUTS is dependent upon the number of tuning samples, which was set to be 25-33% as large as the number of samples retained afterwards.

For prediction on holdout test data, the posterior mean (for NUTS and Gibbs) or variational posterior mean (for ADVI and IB-CAVI) was used as a point estimate $\widehat{\boldsymbol{B}}$ for $\boldsymbol{B}$. This value $\widehat{\boldsymbol{B}}$ was then substituted into the appropriate category probability formula – softmax, multi-logit (i.e. identified softmax), CB-Probit, or CB-Logit. For a given CB link function (probit or logit), the CB variant (CBM or CBC) was chosen that yielded the largest training likelihood. This strategy provides a cheap heuristic approximation to BMA, as most datasets have sufficiently many observations that the BMA weights tend to be very close to 0.0 or 1.0.

For performance metrics, we used mean holdout log-likelihood and mean predictive accuracy. The mean holdout log-likelihood was computed in the standard way (Eq. (G.1.2)). For the accuracy metric, the category with the largest probability was considered to be the predicted category. If a test set observation had C categories predicted with the same probability, then the model was given credit for $1/C$ rather than 1 correct response. For simulated data, we also computed these performance metrics under random guessing and when using the true model (i.e. softmax regression, using $\boldsymbol{B}_{\text{true}}$.)

G.4.4 Results

The primary results were given in Sec. 5.3. Supplemental results are provided in Fig. G.2.

G.5 The impact of the IB-approximation on posterior over category probabilities

In this section, we directly investigate the quality of the posterior over category probabilities that is learned by IB-CAVI. By *posterior over category probabilities*, we refer to the categorical likelihoods $p(y = k|\boldsymbol{B})$ obtained by drawing the regression weights from the approximate posterior density over weights $q(\boldsymbol{B} \mid \boldsymbol{y}_{1:N})$, where $\boldsymbol{y}_{1:N}$ is the training data. While a direct analysis of $q(\boldsymbol{B} \mid \boldsymbol{y}_{1:N})$ is possible, this is an intermediate quantity less relevant to applications (see Sec. 1.1) and may be confounded by identifiability issues.

Thus, we compare IB-CAVI's posterior over category probabilities against that learned by other methods that do not make an IB-approximation. We would like to obtain a concrete visualization of how the IB-approximation impacts the bias and variance of this posterior over category probabilities. Of particular interest is the comparison to the NUTS sampler, which can be taken as the gold standard.

G.5.1 Methodology

Dataset. We construct a simulated dataset using the method described in Sec. G.1.1 with N=1000 samples, K=4 categories, and M=8 covariates. We set $\sigma^2_{\text{high}} = 4.0$.

Methods. We train a CB-Probit model with IB-CAVI (Algorithm 2) until the drop in the mean ELBO (with the mean taken across the number of samples N and categories K) was less than 0.01 across consecutive iterations. Bayesian model averaging (BMA; Sec. 4) reveals that the weight on the CBC model, π_{CBC}, was very close to 1.0; thus, the predictions of the CB-Probit model with BMA is virtually identical to the predictions of the CBC-Probit model. For this reason, our baseline black-box inference methods use the CBC-Probit (rather than CBM-Probit, or some mixture). In this case, the baseline methods used were Automatic Differentiation Variational Inference (ADVI) or the No-U-Turn Sampler (NUTS) (see Sec. G.4.2).

G.5.2 Results

Fig. G.3 gives the posterior over category probabilities for the first 9 training set observations as approximated by IB-CAVI, ADVI, and NUTS.

G.5.3 Discussion

IB-CAVI delivers posteriors over category probabilities that are reasonably good approximations to those obtained by ADVI and NUTS. However, the procedure does appear to reduce variance and introduce some bias. For applications where fidelity to the true posterior is critical, one could use IB-CAVI for warm-starting. That is, one could use IB-CAVI's quickly learned approximate posterior to initialize a more expensive procedure that delivers greater fidelity. For example, one might use IB-CAVI to initialize NUTS, which is computationally expensive but asymptotically exact.

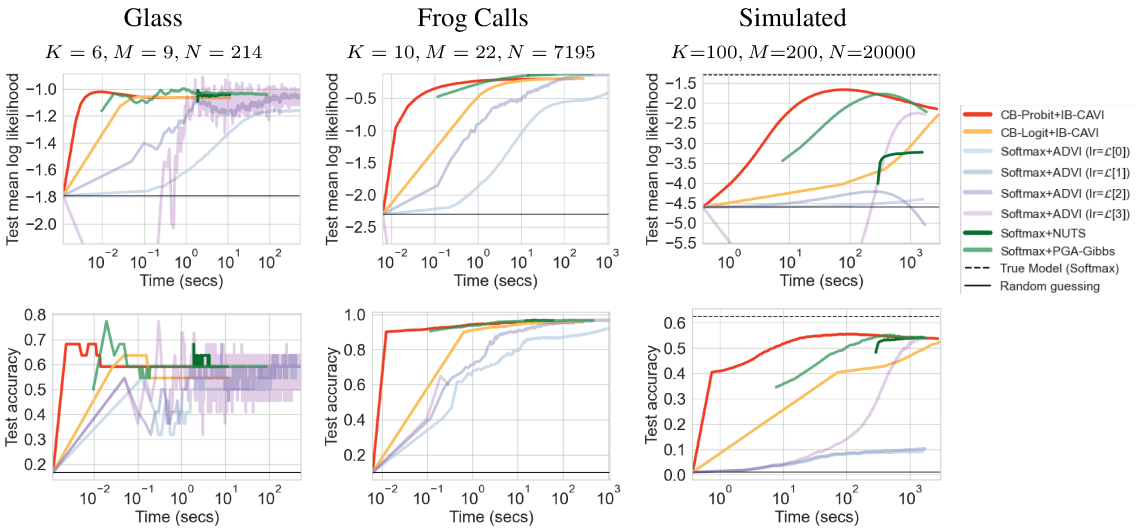

Figure G.2: Supplemental comparisons of holdout log likelihood (top) and accuracy (bottom) over training time on real and simulated datasets with K categories, M covariates, and N instances. For ADVI, we try the learning rates $\mathcal{L} := (0.01, 0.1, 1.0, 10, 100)$ recommended by (Kucukelbir et al., 2017), adjusted to $10^{-1}\mathcal{L}$ in the largest dataset to reduce divergence. If a line is absent for ADVI, the method diverged. Note that for IB-CAVI, the parallelism over K could be exploited to yet further reduce training time.

G.6 Intrusion detection experiment: Supplemental information

G.6.1 Data

The raw multi-source cyber-security behavioral event data (Kent, 2015) contains 58 consecutive days of computer usage behavior from within Los Alamos National Laboratory's corporate, internal computer network. Behavior is represented in terms of events from five modalities (process starts, network activity, etc.) For earlier work on intrusion detection with this dataset, see (Turcotte et al., 2016).

We restrict our analysis to the raw process data, which represents process start and stop events collected from individual Windows-based desktop computers and servers. In addition, we restrict our attention to process starts from human users on active directory domain accounts. This restriction requires three pre-processing steps:

1. Use only the process starts (discard the process ends).
2. Restrict to human users. Users with names beginning with a 'U' are human accounts while those beginning with a 'C' are computer accounts (managed by computers not actual people). Correspondingly, users which start with 'C' seemed to have less variation across users in process starts.
3. Restrict to active directory domain accounts. Domains starting with 'C' are local accounts typically used on individual workstations. Domains starting with 'DOM' can be used to authenticate on local machines, but they are also the standard way of authenticating for network resources (email, databases, servers, etc.). There is a lot more user overlap for 'DOM' resources, making them less predictable.[11]

G.6.2 Featurization

During an exploratory data analysis, we notice:

1. Regularities in process start subsequences can have minor permutations. For instance, see user U1788, who deterministically cycles through P111-P296-P298-P299, until breaking out of this pattern for the last 10 or so process starts. The same processes are used, but in a different pattern.
2. Multiple processes can be launched simultaneously (up to the single second resolution with which time is reported), and the order in which simulataneous processes are listed is not invariant. (Note this provides a partial explanation for item (1).)

To accommodate these features of the data, we do not use a strict autoregressive featurization, but a softer version which should be more tolerant of noise. In particular, we

[11] This information was provided by personal communication with Aaron Scott Pope, the current contact from Los Alamos National Laboratories provided by https://csr.lanl.gov/data/cyber1/.

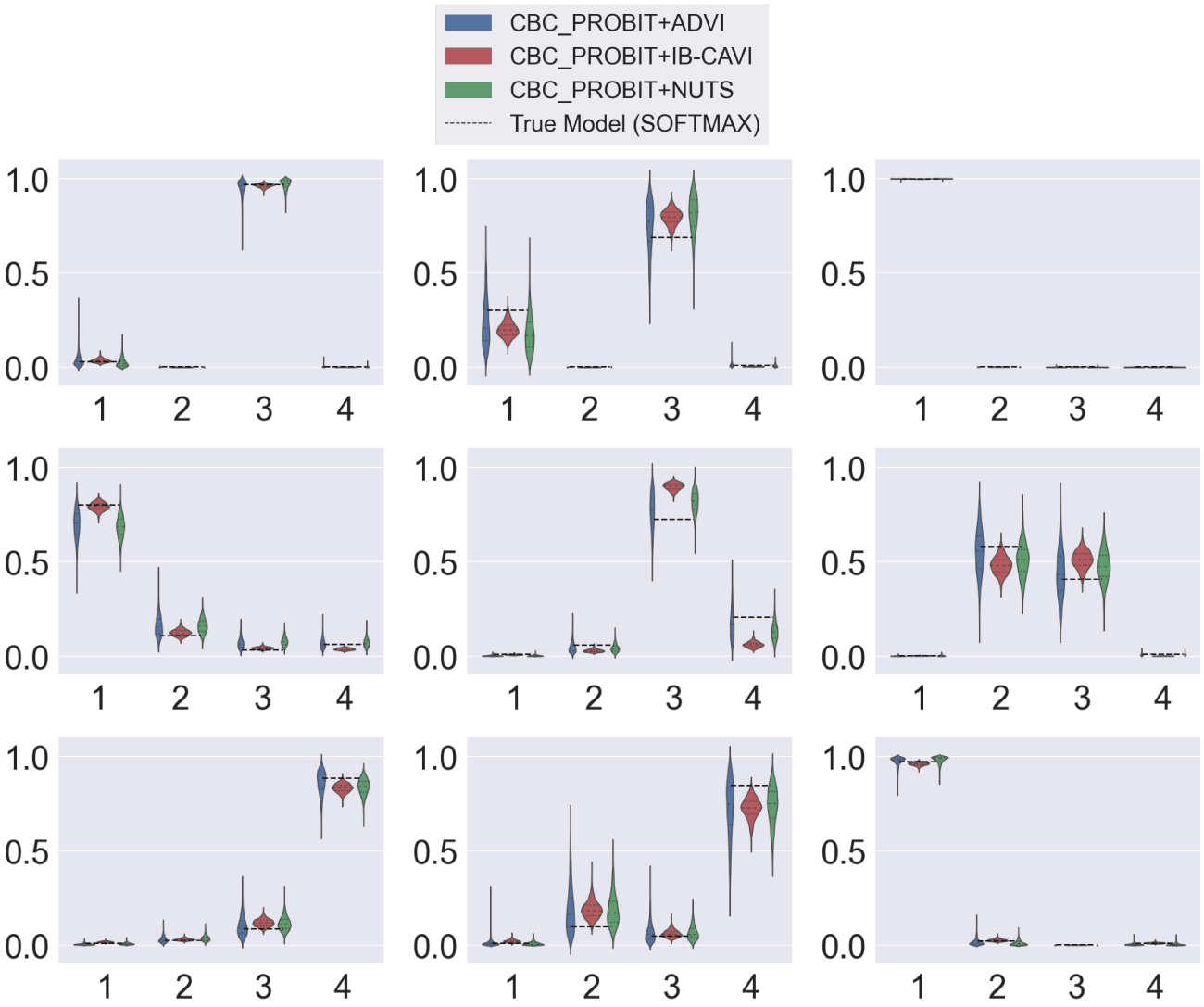

Figure G.3: Violin plots showing the approximate posteriors over category probabilities (y-axis) across four possible categories (x-axis). The approximate posteriors over category probabilities are given by three approximate Bayesian inference methods applied to simulated softmax regression data. Each subplot represents the approximate posterior predictive distribution over categories given the covariates for a single training set observation.

choose a lookback window of W processes, and then featurize each of these W processes with $\exp(-\Delta t/\tau)$ seconds, where τ is a temperature parameter and Δt refers to how long ago (in seconds) the process was launched.

The window size W could perhaps be justified by plotting the distribution on the number of simultaneous process launches, and saying that the window size is the whatever-th percentile of that distribution.

G.6.3 Methodology for experiment

Here we describe the methodology used for the experiment discussed in Sec 5.5. We selected U=32 users from the database who had moderately many process starts. The number of processes started per user, N_u, over the course of the 58 days of data collection ranged from 17,678 to 19,261.[12]

We learn each user's process start behavior by training a separate model for each user. We use the featurization strategy of Section G.6.2, somewhat arbitrarily choosing the window size to be $W = 5$ and the temperature to be $\tau = 60$ seconds. We take the number of categories to be $K = 1,553$, the number of unique processes in the entire dataset.

We take the first 80% of the process start events to be training data, and the remainder to be hold-out test data. We use IB-CAVI to approximately learn the CBC-Probit model. We ran inference for 100 iterations. Each iteration required approximately 5 to 20 seconds of computation time.

G.7 Glass identification: Supplemental analysis

Here we provide further analysis of the glass identification dataset that was also analyzed in the holdout performance over time experiment (Sec. G.4). Here, following (Johndrow et al., 2013), we perform 10-fold cross validation, randomly splitting the dataset 10 times into a training set and test set, where each split put 90% of the original dataset into the training set. Thus, each data split had $N_{\text{train}} = 192$ training samples, and $N_{\text{test}} = 22$ test samples.

We z-transformed all variables, as the range of some variables is very small (e.g. consider the `RI` variable, which only varies from 1.51 to 1.52). This lets us use independent $\mathcal{N}(0, 1)$ priors on the regression weights for each covariate-category combination.

G.7.1 Methodology

We applied two different Bayesian inference methods : MCMC sampling and variational inference. For MCMC sampling, we applied the implementation of the No U-Turn Sampler (NUTS) (Hoffman et al., 2014) given in the `numpyro` library. We obtained 10,000 total samples (3,000 burn-in samples). For variational inference, we applied IB-CAVI, and concluded convergence when the drop in the mean ELBO (with the mean taken across the number of samples N and categories K) was less than 0.005 across consecutive iterations. For both inference methods, we initialized the regression weights $\boldsymbol{B}$ to the zero matrix.

G.7.2 Results

Table G.3 shows the results. We find that IB-CAVI gives results that are close to those obtained by NUTS, but between 44 and 1,110 times faster. We also note from the NUTS results that the CB models perform competitively with the softmax model, a much more familiar categorical GLM.

[12] The target number N_u serving as an inclusion criterion was chosen out of convenience: the Python package in its currently implementation can handle $N_u \approx 20,000$ without a memory error, but cannot handle the largest value of N_u in the dataset, due to memory constraints. No attempt was made to model the largest N_u, because the experiment as is seems sufficient to prove the point. Further scalability could be obtained by improving the implementation (in terms of handling of sparsity and/or further exploiting the fact that the algorithm is embarassingly parallel across categories), or by incorporating memoization (Hughes & Sudderth, 2013) or stochastic variational inference (Hoffman et al., 2013) strategies within the IB-CAVI framework.

Table G.3: *Glass identification results.* The (geometric) mean holdout likelihood is given by $\exp\left(\frac{1}{FN_{\text{test}}}\sum_{f=1}^{F}\sum_{n=1}^{N_{\text{test}}}\log p(y_n^{\text{test}} \mid \boldsymbol{B}^*)\right)$, where F is the number of cross-validation folds and $\boldsymbol{B}^*$ is the posterior expectation from IB-CAVI or NUTS. It represents the typical probability score that the fitted model assigns to categorical outcomes in the test set. Computation time is measured in seconds. Note that accuracy will always be identical for CBC and CBM models with the same IB base model when fit with IB-CAVI, as guaranteed by Prop. B.5.1.

Model	Softmax	CBC-Logit		CBM-Logit		CBC-Probit		CBM-Probit	
Inference	NUTS	NUTS	IB-CAVI	NUTS	IB-CAVI	NUTS	IB-CAVI	NUTS	IB-CAVI
Mean likelihood	0.38	0.38	0.36	0.38	0.36	0.34	0.35	0.41	0.37
Accuracy	0.64	0.65	0.64	0.64	0.64	0.65	0.65	0.64	0.65
Computation time	20.17	26.07	0.30	15.41	0.30	333.04	0.35	59.98	0.35