Easy Variational Inference for Categorical Observations via a New View of Diagonal Orthant Probit Models

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

[1]Data Intensive Studies Center, Tufts University, Medford, MA, USA
[2]Dept. of Electrical and Computer Engineering, Tufts University, Medford, MA, USA
[3]Dept. of Computer Science, Tufts University, Medford, MA, USA

Abstract

In pursuit of tractable Bayesian analysis of categorical data, auxiliary variable methods hold promise, but impose asymmetries on the truly unordered categories or spoil scalability via strong dependencies in posteriors over parameters. The Diagonal Orthant Probit (DO-Probit) model proposed by Johndrow, Lum, and Dunson (AISTATS 2013) avoids these difficulties, treating all categories symmetrically while yielding tractable conditionally conjugate inference. However, we show that the intended DO-Probit likelihood for categorical observations, when paired with a normal prior, does not yield a conjugate posterior. Instead, we clarify that their posterior analysis is only correct for a different model that treats observations as multiple independent binary draws. This raises two questions: Other than tractability, what justifies the binary model for categorical data? And how should a binary model make categorical predictions? To resolve these issues, using variational methods we obtain a lower bound of a categorical model's marginal likelihood that can be optimized by fitting the conjugate binary model. Optimizing this bound enjoys all benefits advocated in the original DO-Probit work. We further extend this fast, reliable covariate-informed modeling of categorical outcomes to groups or sequences of data related in a hierarchy.

1 INTRODUCTION

We consider the problem of modeling categorical data informed by covariates using the machinery of generalized linear models. Because our intended big data applications may involve rare events or little available data for some quantities of interest, we pursue Bayesian analysis 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 observations have been tried, Bayesian analysis of these models remains a difficult problem 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 model is not conjugate, and so estimating posteriors over weight parameters requires expensive sampling methods [Hoffman and Gelman, 2014] or non-conjugate variational optimization methods [Wang and Blei, 2013, Braun and McAuliffe, 2010, Kucukelbir et al., 2017]. Recent auxiliary variable methods [Polson et al., 2013] have yielded expanded binary logistic models with conjugate conditionals, but extensions to 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 and 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 and Chib, 1993, Held and Holmes, 2006]. However, multi-class probit models encode strong posterior dependence between 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's strong dependence results in a lack of closed-form category probabilities [Johndrow et al., 2013], which alone has prevented adoption within more complicated models (e.g., see [Holsclaw et al., 2017]).

Accepted paper for the 4th Workshop on Tractable Probabilistic Modeling (TPM 2021).

Motivated by difficulties that arise from these previous efforts, Johndrow et al. [2013] introduced the diagonal orthant multinomial probit (DO-Probit) model. Their proposed construction treats all categories symmetrically, yields tractable category probabilities, and achieves conditionally conjugate posteriors while avoiding any cross-category dependence among auxiliary variables. These features should make the DO-Probit a prime candidate for fast, scalable Bayesian modeling of categorical data.

However, we have uncovered subtle but critical gaps in the technical justification for the Bayesian analysis of the DO-Probit provided by Johndrow et al. [2013]. In particular, we find that the DO-Probit categorical likelihood, when paired with a normal prior, does not admit conditionally-conjugate posteriors even with auxiliary variables. The Gibbs sampler presented in Sec. 3 of Johndrow et al. [2013] is not a correct sampler when the intended posterior is formed from a categorical likelihood. Instead, it is a correct sampler only when understood as fitting a *different model*: the independent binary model, which is conditionally conjugate after augmentation via the classic arguments for the binary probit [Albert and Chib, 1993]. This "model swap" may not be obvious to readers of the original DO-Probit text and has apparently gone unnoticed in some later uses of this method. Magnusson et al. [2020] use the Gibbs sampler of Johndrow et al. [2013] directly for a categorical likelihood, which is not a valid posterior sampling technique.

The need for a model swap raises important questions: given categorical data, what justifies using the independent binary model, which is not intended for one-of-K categories, other than tractability? If we insist on using the binary model, how can we use it to make valid categorical predictions?

This paper makes several key contributions in attempting to answer these questions:

1. We show that the DO-Probit model for categorical data does *not* have a conditionally-conjugate posterior over weights even with auxiliary variables (Sec. 2).
2. We suggest an effective way to make predictions about heldout categorical data using the independent binary model (Sec. 2.6), which in the intercepts-only case surprisingly achieves indistinguishable prediction quality compared to models designed for categorical data.
3. We introduce a new generalized linear model for categorical data, which we call the Simplified DO-Probit model (SDO-Probit, Sec. 2). This model has similar benefits as the original DO-Probit (symmetric treatment of categories and closed-form category probabilities computed using the probit function), but yields the effective prediction strategy naturally.
4. We derive a tractable lower bound for the marginal likelihood of the Simplified DO-Probit categorical model, and show that optimizing this bound corresponds to fitting the observed data using the independent binary model (Sec. 3). This allows us to *justify* applying the conjugate machinery suggested by Johndrow et al. [2013] to otherwise intractable categorical models. Our bound holds for the original DO-Probit model as well.
5. Using our bound, we provide an easily-implemented variational inference strategy for categorical SDO-Probit model that is fast and reliable. In particular, the inference has three beneficial properties: (a) closed-form updates, (b) decoupled auxiliary variables, and (c) symmetric handling of categories. No prior work known to us achieves all three properties. We provide an algorithm which allows the technique to be extended to *any* model which uses the SDO- (or DO-) Probit complete likelihood (see Algorithm 1 in Sec. 3.2).
6. We highlight the advantages of this flexibility by introducing a *hierarchical* DO-Probit model for modeling groups or sequences of categorical data (Sec. 4).

In summary, our new variational methods provide a principled route to fast, scalable modeling of categorical data using a surrogate binary model.

2 MODELS

We now formally introduce the models of interest and establish their properties. First, we introduce one model for observed categorical data, the diagonal orthant probit, denoted throughout as DO, first described in Johndrow et al. [2013]. Next, we introduce an alternative model for multivariate binary data, which we refer to as "independent binary" or IB, first described in Albert and Chib [1993]. Finally, we introduce a new model for categorical data, which we call the Simplified-DO Probit (SDO). We briefly discuss maximum likelihood estimation for each model. We then expose that in a posterior estimation context, neither categorical model is conjugate even with auxiliary variables, due to an intractable normalization constant. Thus, the Gibbs sampler presented by Johndrow et al. [2013] for the "DO-Probit" model is in fact only a valid sampler for the IB. This raises the question of *why* one should apply the IB to categorical data, which we address in Sec. 3.

2.1 DO-PROBIT MODEL FOR CATEGORY DATA

The *diagonal orthant probit* model [Johndrow et al., 2013] describes the generative process for N categorical outcomes, where each observation (indexed by $i \in \{1, \dots N\}$) is a one-of-K indicator $y_i \in \{1, 2, \dots K\}$. These outcomes are generated by transforming known covariate vector $\boldsymbol{x}_i \in \mathbb{R}^M$ with unknown weight parameters $\boldsymbol{\beta} \in \mathbb{R}^{M \times K}$. Let $\boldsymbol{\beta}_k$ designate the k-th column of $\boldsymbol{\beta}$.

To generate observation i, we first draw a latent continuous vector $\boldsymbol{z}_i \in \Omega \subset \mathbb{R}^K$ from a multivariate normal whose mean is determined by the dot product of covariates and weights but whose possible values are *truncated* to the Diagonal Orthant region Ω. This region con-

tains all K-dimensional vectors with one entry positive and the rest negative. Formally, $\Omega := \bigcup_{k=1}^{K} \Omega_k$, where $\Omega_k = \{\boldsymbol{\omega} \in \mathbb{R}^K : \omega_k \geq 0, \omega_j < 0 \; \forall \; j \neq k\}$.

Given vector $\boldsymbol{z}_i$, our second generative step is simply to set the outcome y_i to indicate which entry of $\boldsymbol{z}_i$ is positive. This two-step sampling is formalized for each $i \in \{1, \dots N\}$:

$$\boldsymbol{z}_i \mid \boldsymbol{\beta} \overset{\text{ind}}{\sim} \mathcal{TN}\left(\begin{bmatrix} \boldsymbol{x}_i^T \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_K \end{bmatrix}, \boldsymbol{I}_{K \times K}, \Omega \right),$$
$$y_i = k \iff z_{ik} > 0. \tag{2.1.1}$$

Here, $\mathcal{TN}(\boldsymbol{\eta}, \boldsymbol{V}, R)$ defines a truncated multivariate normal distribution formed when a normal with mean $\boldsymbol{\eta}$ and covariance $\boldsymbol{V}$ is truncated to region R.

The DO's complete-data likelihood is:

$$p_{\text{DO}}(y_i, \boldsymbol{z}_i \mid \boldsymbol{\beta}) = \frac{1}{C_i(\boldsymbol{\beta})} 1_{\boldsymbol{z}_i \in \Omega} \cdot \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_{ik} - \boldsymbol{x}_i^T \boldsymbol{\beta}_k)^2 \right) \left(1_{z_{ik} \geq 0}^{1(y_i = k)} 1_{z_{ik} < 0}^{1(y_i \neq k)} \right) \tag{2.1.2}$$

$$\text{where } C_i(\boldsymbol{\beta}) := \sum_{k=1}^{K} \left(\Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}_k) \prod_{j \neq k} \Phi(-\boldsymbol{x}_i^T \boldsymbol{\beta}_j) \right) \tag{2.1.3}$$

and where Φ is the CDF of a standard normal. By construction, $C_i(\boldsymbol{\beta})$ is the fraction of probability mass of the unconstrained Gaussian contained in the set Ω, i.e. the orthants with only one positive entry, and thus $0 < C_i(\boldsymbol{\beta}) < 1$. This fact will be useful in Sec. 3.

We can view $\boldsymbol{z}_i$ as an auxiliary that can be marginalized out, yielding category probabilities conditioned only on $\boldsymbol{\beta}$:

$$p(y_i = k \mid \boldsymbol{\beta}) = p(z_{ik} \geq 0, z_{ij} < 0, j \neq k \mid \boldsymbol{\beta}) \tag{2.1.4}$$
$$= \frac{\Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}_k) \left(\prod_{j \neq k} \Phi(-\boldsymbol{x}_i^T \boldsymbol{\beta}_j) \right)}{\sum_{k=1}^{K} \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}_k) \left(\prod_{j \neq k} \Phi(-\boldsymbol{x}_i^T \boldsymbol{\beta}_j) \right)}$$

This *tractable* formula was first given in Johndrow et al. [2013, p. 4]. We emphasize that the standard multi-class probit model has no such closed form.

2.2 A PROBIT MODEL FOR BINARY VECTORS

We now consider another model, which we call the *Independent Binary* probit model (IB). This model produces N observations (indexed by i), where each observation a binary vector $\bar{\boldsymbol{y}}_i$ of size K. Crucially, $\bar{\boldsymbol{y}}_i$ is *not* a one-hot vector: any number of entries could be 1 or 0. The variables here play similar roles to counterparts in the earlier DO model, but have different domains. Thus, we use symbols with *bars* to denote IB variables.

Following Albert and Chib [1993], to generate observation i using the IB model, we first sample an auxiliary variable $\bar{z}_{ik}$ for each label k from a Normal whose mean is the dot product of covariates $\boldsymbol{x}_i$ and weights $\bar{\boldsymbol{\beta}}_k \in \mathbb{R}^M$:

$$\bar{z}_{ik} \mid \bar{\boldsymbol{\beta}}_k \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}_k, 1). \tag{2.2.1}$$

Second, we use the sign of the value of $\bar{z}_{ik}$ to deterministically set the binary observation for label k:

$$\bar{y}_{ik} = \begin{cases} 1 & \bar{z}_{ik} \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2.2}$$

The complete data likelihood for the IB model is thus:

$$p_{\text{IB}}(\bar{\boldsymbol{y}}_i, \bar{\boldsymbol{z}}_i \mid \bar{\boldsymbol{\beta}}) = \prod_{k=1}^{K} p(\bar{y}_{ik}, \bar{z}_{ik} \mid \bar{\boldsymbol{\beta}}) \tag{2.2.3}$$
$$= \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\bar{z}_{ik} - \boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}_k)^2 \right) \left(1_{\bar{z}_{ik} \geq 0}^{1(\bar{y}_{ik} = 1)} 1_{\bar{z}_{ik} < 0}^{1(\bar{y}_{ik} = 0)} \right).$$

2.3 A SIMPLIFIED DO-PROBIT MODEL

We introduce a third model, which we call the *simplified diagonal orthant probit* or just SDO, which produces categorical observations. Given covariates $\boldsymbol{x}_i$ and weight parameters $\boldsymbol{\beta}$, we produce categorical outcome y_i via:

$$y_i \sim \text{Cat}(p_1, \dots p_K), \quad p_k = \frac{\Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)}{\sum_{\ell=1}^{K} \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}_\ell)}. \tag{2.3.1}$$

This is a simpler way to use the probit link function to produce closed-form categorical probabilities than in the original DO model in Eq. (2.1.4), hence the name "simplified" DO-Probit. While a similar construction appears in Magnusson et al. [2020, p. 6], here for the first time we detail important properties including this model's complete data likelihood with auxiliaries and resulting non-conjugacy.

The SDO is completed by sampling auxiliary variable $\boldsymbol{z}_i \in \Omega_{y_i} \subset \mathbb{R}^K$ from a truncated Normal whose pre-truncated mean is a dot product of weights and covariates like previous models:

$$\boldsymbol{z}_i \mid \boldsymbol{\beta}, y_i \overset{\text{ind}}{\sim} \mathcal{TN}\left(\begin{bmatrix} \boldsymbol{x}_i^T \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{x}_i^T \boldsymbol{\beta}_K \end{bmatrix}, \boldsymbol{I}_{K \times K}, \Omega_{y_i} \right).$$

By definition, $\boldsymbol{z}_i$ has only one positive entry (at index y_i) and all other entries are negative. Thus, this model still uses a diagonal orthant construction.

The SDO complete data likelihood is given by:

$$p_{\text{SDO}}(y_i, \boldsymbol{z}_i \mid \boldsymbol{\beta}) = \frac{1}{C_{i,y_i}^{\text{SDO}}(\boldsymbol{\beta})} 1_{\boldsymbol{z}_i \in \Omega} \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_{ik} - \boldsymbol{x}_i^T \boldsymbol{\beta}_k)^2 \right) \cdot \left(1_{z_{ik} \geq 0}^{1(y_i = k)} 1_{z_{ik} < 0}^{1(y_i \neq k)} \right) \tag{2.3.2}$$

Here, the product of the indicators is sufficient to restrict $z_i \in \Omega_{y_i}$ as required. C^{SDO} is a normalization term,

$$C_{i,k}^{\text{SDO}}(\boldsymbol{\beta}) := \left(\textstyle\sum_{\ell=1}^{K} \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta}_\ell)\right)\left(\textstyle\prod_{j \neq k} \Phi(-\boldsymbol{x}_i^T \boldsymbol{\beta}_j)\right), \tag{2.3.3}$$

which interestingly for this model *depends* on the chosen category index $y_i = k$. Like the previous normalization term, C^{SDO} always produces values between 0 and 1.

While this SDO model enjoys several properties shared by the original DO model (closed-form category probabilities

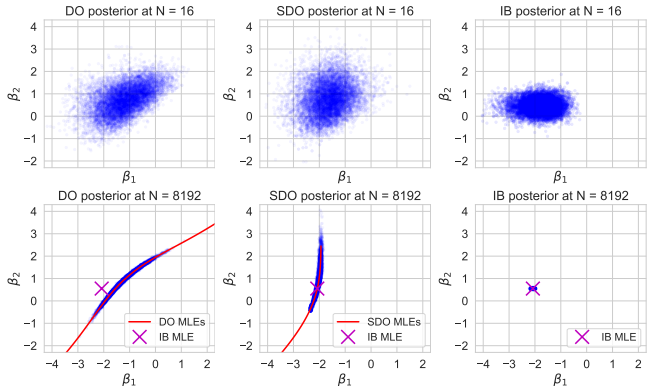


Figure 1: **Comparison of posteriors and ML point estimates for two weight coefficients (β_1, β_2) in *intercepts-only* setting.** *Left:* DO model for categorical observations. *Center:* Simplified DO (SDO) model, also categorical. *Right:* IB model for binary observations. All models are intercept-only (no covariates) with standard Normal priors, and are fit to the first N examples of the same dataset (top row $N = 16$, bottom $N = 8192$), generated with $K = 3$ categories with ground truth frequencies [0.02, 0.7, 0.28].

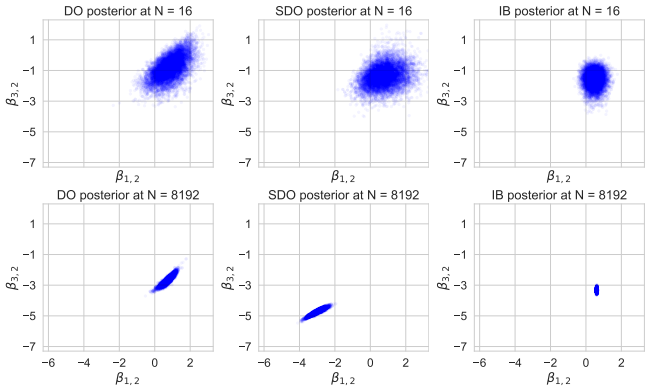


Figure 2: **Comparison of posteriors for two weight coefficients $(\beta_{1,2}, \beta_{3,2})$ fit to synthetic data where each example has *one covariate* drawn from a standard Normal.** *Left:* DO model for categorical observations. *Center:* Simplified DO (SDO) model. *Right:* IB model for binary observations. All models have standard Normal priors and $K = 3$ categories. We fit to the first N examples of the same dataset. Unlike the intercepts-only case, IB and SDO *do not* concentrate on the same weight values as $N \to \infty$.

and symmetric treatment of categories), SDO is particularly useful for connecting the IB model to categorical outcomes.

2.4 POINT ESTIMATION AND IDENTIFIABILITY

In general, there are infinitely many possible $\boldsymbol{\beta}$ parameters that would yield the same category probabilities under the DO model in Eq. (2.1.4). Thus, a maximum likelihood estimate (MLE) for $\boldsymbol{\beta}$ under the DO likelihood is not unique. Similarly, the SDO -likelihood also has a non-unique MLE. Below, we outline key properties about MLE estimators for all models in the *intercepts-only* (no covariates) setting.

Unique MLE for intercept-only IB. Let $p_1, \dots p_K$ be empirical category frequencies in a K-class intercepts-only dataset: $p_k = \frac{1}{N} \sum_{i=1}^{N} 1_{y_i = k}$. As Johndrow et al. [2013] argue, if we encode the observations y_i as one-hot vectors, there is a *unique* maximum likelihood estimate for $\bar{\boldsymbol{\beta}}$ under the IB model, available by setting the weight for each k as:

$$\bar{\beta}_k^* = \Phi^{-1}(p_k). \tag{2.4.1}$$

MLE for intercept-only SDO. A similar argument can be used to define an ML estimate for the SDO in an intercepts-only setting. Suppose we restrict $\sum_{k=1}^{K} \Phi(\beta_k) = r$ for some $r \in (0, \min_\ell \frac{1}{p_\ell})$. Fixing r, the unique MLE for SDO is:

$$\beta_k^* = \Phi^{-1}(r p_k). \tag{2.4.2}$$

When $r = 1$, the SDO MLE for $\boldsymbol{\beta}$ here is *exactly equal* to the IB MLE from Eq. (2.4.1).

MLE for intercept-only DO. Similarly, in an intercepts-only version of the DO model, if we enforce the constraint $\sum_{k=1}^{K} \frac{\Phi(\beta_k)}{\Phi(-\beta_k)} = s$ for a specific $s > 0$, then we can compute the MLE for each category k as:

$$\beta_k^* = \Phi^{-1}\left(\frac{s p_k}{1 + s p_k}\right). \tag{2.4.3}$$

Plugging this value into Eq. (2.1.4) yields category probabilities that *exactly match* the observed frequencies p.

Stepping back, we emphasize that Johndrow et al. [2013]'s only technical justification for using the IB model seems to be that in an *intercepts-only* setting, the IB has unique MLE weights that exactly match the MLE for a categorical DO model[1]. We find this argument for using IB insufficient, as it applies only to the maximum likelihood point estimation setting without covariates. This leaves a critical gap: we lack justification for the IB as a model for categorical data *with covariates* when we seek *posteriors*, not just point estimates.

2.5 BAYESIAN POSTERIOR ANALYSIS

The three models discussed here – DO, SDO, and IB – each induce a *distinct* posterior over weights when given the same data and same prior. Fig. 1 shows each model's posterior over two weight coefficients given toy categorical data in the intercepts-only setting. Fig. 2 does the same given a synthetic dataset *with covariates*. Posteriors are estimated via the NUTS sampler [Hoffman and Gelman, 2014], as implemented in NumPyro [Phan et al., 2019] using JAX for automatic differentiation [Bradbury et al., 2018]. For large N, we also show the unique MLE for IB and the manifold of possible (non-unique) ML estimates of weights for DO and

[1] Johndrow et al. motivate the IB via a "marginal MLE restriction", which we think corresponds to the special case where IB's MLE exactly matches *SDO*'s MLE when $r = 1$. No such MLE relation exists for DO. Johndrow et al. do not define the SDO explicitly, however.

SDO, using formulas in Eq. (2.4.1)-(2.4.3) (these formulas only apply in Fig. 1).

In the no-covariate case of Fig. 1, SDO and IB share an MLE, and thus for large N the IB posterior concentrates its mass where the SDO does. However, in Fig. 2, all three models have distinct posteriors and even with $N = 8192$ examples, the IB posterior does not necessarily concentrate where either DO or SDO do.

We now assess the conjugacy properties of each model. We assume the weight parameters are given a normal prior.

Observation 1: The IB model's complete conditionals all have conjugate form. Classic arguments about the binary probit [Albert and Chib, 1993, Held and Holmes, 2006] show that the intended posterior conditionals can be written as a Normal for $p_{\text{IB}}(\bar{\boldsymbol{\beta}} \mid \bar{\boldsymbol{z}}, \bar{\boldsymbol{y}})$ and a truncated Normal for $p_{\text{IB}}(\bar{\boldsymbol{z}} \mid \bar{\boldsymbol{\beta}}, \bar{\boldsymbol{y}})$. This posterior is clearly the one fit by the Gibbs sampler in Johndrow et al. [2013].

Observation 2: The DO model's complete conditional for weights $\boldsymbol{\beta}$ is not normal and thus not conjugate. Combining a normal prior over $\boldsymbol{\beta}$ with the DO-Probit likelihood in Eq. (2.1.2), the resulting posterior $p(\boldsymbol{\beta} \mid \boldsymbol{z}, \boldsymbol{y})$ includes the normalizing constant $C_i(\boldsymbol{\beta})$ in Eq. (2.1.3), whose functional form spoils conjugacy: including this term means the log posterior is not a quadratic function of $\boldsymbol{\beta}$.

Observation 3: The SDO model's complete conditional for weights $\boldsymbol{\beta}$ is not normal and thus not conjugate. We follow a similar argument as the DO model. The complete data likelihood of Eq. (2.3.2) contains the normalizing constant $C^{\text{SDO}}_{i,y_i}(\boldsymbol{\beta})$ (Eq. (2.3.3)) that spoils conjugacy.

Thus, both DO and SDO categorical models are not conjugate, simply by inspection of the complete data likelihood. This is not stated clearly in Johndrow et al. [2013] and has led others to assume their Gibbs sampler is correct for a categorical likelihood, which is not true.

So while Johndrow et al. [2013] advocate for Bayesian analysis of categorical data with DO Probit models, in fact the Gibbs sampler they derive is only correct for the independent binary generative model. This leaves two gaps in understanding: How should the IB make predictions for categorical data? More importantly, why is the IB appropriate for Bayesian analysis of categorical data? We answer the first question in Sec. 2.6, and the second in Sec. 3.

2.6 PREDICTIONS FOR CATEGORICAL DATA

Suppose we have specific weight values $\bar{\boldsymbol{\beta}}$ for the IB that have been fit to purely categorical data. For example, this may be an ML estimate of the weights or a posterior sample. We wish to predict a new observation, which we are reliably informed will be categorical. The IB's *support* is then misspecified, because it could produce any binary vector, not just those with one positive entry.

IB+DO. It seems natural that a new observation (indexed by $*$) belongs to category k with probability proportional to the IB likelihood producing the one-hot vector e_k:

$$\hat{p}_{+DO}(y_* = k|\boldsymbol{x}_*, \bar{\boldsymbol{\beta}}) = \frac{p_{\text{IB}}(\bar{\boldsymbol{y}}_* = e_k|\bar{\boldsymbol{\beta}}, \boldsymbol{x}_*)}{\sum_{\ell=1}^{K} p_{\text{IB}}(\bar{\boldsymbol{y}}_* = e_\ell|\bar{\boldsymbol{\beta}}, \boldsymbol{x}_*)} \tag{2.6.1}$$

Here, we normalize over the set of possible one-hot vectors of size K. We call this the IB+DO estimator of heldout likelihood, because mathematically this reduces to plugging IB weights $\bar{\boldsymbol{\beta}}$ into DO's probability formula in Eq. (2.1.4),

IB+SDO. An alternative estimator of heldout likelihood is:

$$\hat{p}_{+SDO}(y_* = k|\boldsymbol{x}_*, \bar{\boldsymbol{\beta}}) = \frac{\Phi(\boldsymbol{x}_*^T \bar{\boldsymbol{\beta}}_k)}{\sum_{\ell=1}^{K} \Phi(\boldsymbol{x}_*^T \bar{\boldsymbol{\beta}}_k)}, \tag{2.6.2}$$

which we refer to as "IB+SDO", because it uses our Simplified DO formula for translating the IB weights $\bar{\boldsymbol{\beta}}$ into category probabilities. For another motivation, see Sec. B.

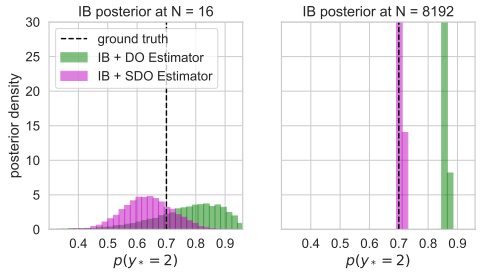
These two estimators are compared both visually and quantitatively in Fig. 3, where posterior samples of weights from the same IB model are plugged into each estimator to predict heldout categorical data. We see significant practical reasons to prefer the IB+SDO estimator in this intercepts-only setting, as it matches the true category frequencies far better than IB+DO. Despite the fact that the IB model is misspecified (it is not naturally a model for categorical data), by fitting the IB model and then applying the SDO estimator (which ensures valid categorical predictions) we can surprisingly deliver heldout likelihoods that for all dataset sizes $N > 16$ are *indistinguishable* from models like DO or SDO that directly capture the categorical nature of data.

3 VARIATIONAL METHODS

We have clearly established that [Johndrow et al., 2013]'s tractable posterior is targeted at the independent binary model, which is not a model for categorical data. In this section, we provide a principled justification for applying a binary model to categorical data, based on relating the binary model and our suggested Simplified DO-Probit categorical model via variational methods.

Variational inference [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".

Our strategy is to show that this lower bound under the misspecified (but tractable, conjugate) IB model, is in turn a lower bound for the same data under the desired (but intractable, non-conjugate) SDO model. Thus, by using the conjugate machinery suggested in Johndrow et al. [2013] to obtain closed-form coordinate ascent updates that maximize this surrogate bound, we are not merely swapping in a convenient model that undesirably is unaware of the categorical nature of our data, but instead exactly improving the fit of



N	IB + DO	IB + SDO	SDO	DO	Cat. MLE
2	-1.6923	-0.9221	-0.9754	-0.8869	-5.6252
16	-1.3300	-0.7045	-0.7251	-0.6958	-5.5333
128	-1.3089	-0.6902	-0.6908	-0.6907	-0.6970
1024	-1.3067	-0.6885	-0.6885	-0.6885	-0.6886
8192	-1.3070	-0.6887	-0.6887	-0.6887	-0.6887

Figure 3: Comparison of categorical predictions using posterior samples from the IB binary model. In this intercept-only (no covariates) setting, we find that IB + SDO offers indistinguishable performance from (harder-to-fit, non-conjugate) categorical models, and superior performance to the IB+DO alternative. IB generates binary vectors by design and thus requires a post-hoc estimator (denoted "+DO" or "+SDO"; defined in Sec. 2.6) to ensure valid categorical predictions. *Left:* Visualization of posterior over the probability of category 2, whose true frequency is 0.7. *Right:* Average log likelihood on heldout set for each model as size of training set N increases. For most methods, we compute the predictive likelihood $\log p(y_*|y_{1:N})$, marginalizing out weights via Monte Carlo averages of posterior samples. Cat MLE baseline just point estimates category frequencies via maximize likelihood. *Setting:* We fit to the first N examples of a toy dataset, generated with $K = 3$ categories with ground truth frequencies [0.02, 0.7, 0.28]. All models use a standard Normal prior over weights.

a suitable generative model. Our strategy will later allow us to easily extend the SDO to more complicated graphical models, as we illustrate in Sec. 4.

3.1 RELATING MODELS VIA BOUNDS ON MARGINAL LIKELIHOODS

We begin with a variational treatment of the SDO model (Eq. (2.3.2)) with a Normal prior density π on $\boldsymbol{\beta}$. The evidence of interest is $p_{\text{SDO}}(\boldsymbol{y}) = \int\int p_{\text{SDO}}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})d\boldsymbol{z}d\boldsymbol{\beta}$. If we select Q as any distribution over $(\boldsymbol{z} \in \Omega, \boldsymbol{\beta} \in \mathbb{R}^{M\times K})$, and let $q(\cdot)$ be the density of Q, then the traditional lower bound of the log of the evidence, which we denote $\texttt{ELBO}_{\text{SDO}} \leq \log p_{SDO}(\boldsymbol{y})$, follows from Jensen's inequality:

$$\texttt{ELBO}_{\text{SDO}}(q) = \underbrace{\mathbb{E}_q[\log p_{\text{SDO}}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta})\pi(\boldsymbol{\beta})]}_{\text{energy}} \underbrace{-\mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{\beta})]}_{\text{entropy}}.$$

Unfortunately, this lower bound is intractable due to the SDO normalizing constant (Eq. (2.3.3)); we must handle N terms of the form $\mathbb{E}_q[\log C^{\text{SDO}}_{i,y_i}(\boldsymbol{\beta})]$, which lack closed-form expression and thus require high-dimensional integral approximations when the number of categories is large.

However, we observe that given the same data y_i and a corresponding auxiliary variable $\boldsymbol{z}_i \in \Omega_{y_i}$, the SDO model will always assign a higher likelihood than the IB:

$$p_{\text{SDO}}(y_i, \boldsymbol{z}_i \mid \boldsymbol{\beta}) > p_{\text{IB}}(\bar{\boldsymbol{y}}_i = e_{y_i}, \bar{\boldsymbol{z}}_i = \boldsymbol{z}_i \mid \bar{\boldsymbol{\beta}} = \boldsymbol{\beta}) \tag{3.1.1}$$

where e_{y_i} is the one-hot vector where index $y_i \in \{1, \dots K\}$ is positive. The relation in Eq. (3.1.1) follows immediately from the complete data likelihoods (2.3.2) and (2.2.3), along with the observation that $C^{\text{SDO}}_{i,y_i}(\boldsymbol{\beta})$ produces values between 0 and 1. It is always valid to provide values from the domain of the SDO model to the IB model.

Next, we argue by monotonicity of the integral, that the *expected complete likelihood* of SDO is bounded by the corresponding expectation under IB for all observations i:

$$\mathbb{E}_q[\log p_{\text{SDO}}(y_i, \boldsymbol{z}_i \mid \boldsymbol{\beta})] > \mathbb{E}_q[\log p_{\text{IB}}(\bar{\boldsymbol{y}}_i{=}e_{y_i}, \bar{\boldsymbol{z}}_i{=}\boldsymbol{z}_i \mid \bar{\boldsymbol{\beta}}{=}\boldsymbol{\beta})]. \tag{3.1.2}$$

This left hand side is a critical additive piece of the SDO ELBO, and the right hand side is a tractable surrogate bound with no troublesome normalizing constant C^{SDO}.

We therefore define a surrogate objective $\mathcal{L}_{\text{SDO}}(q)$ that lower bounds SDO's log marginal likelihood (Eq. (2.3.2)):

$$\begin{aligned}
&\log p_{\text{SDO}}(\boldsymbol{y}) \geq \texttt{ELBO}_{\text{SDO}}(q; \boldsymbol{y}) \\
&= \mathbb{E}_q[\log p_{\text{SDO}}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta})] + \mathbb{E}_q[\log \pi(\boldsymbol{\beta})] - \mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{\beta})] \\
&\overset{(3.1.2)}{>} \mathbb{E}_q[\log p_{\text{IB}}(\bar{\boldsymbol{y}}, \bar{\boldsymbol{z}} \mid \bar{\boldsymbol{\beta}})] + \mathbb{E}_q[\log \pi(\bar{\boldsymbol{\beta}})] - \mathbb{E}_q[\log q(\bar{\boldsymbol{z}}, \bar{\boldsymbol{\beta}})] \\
&= \texttt{ELBO}_{\text{IB}}(q; \bar{\boldsymbol{y}}) := \mathcal{L}_{\text{SDO}}(q)
\end{aligned} \tag{3.1.3}$$

We call this a *surrogate lower bound* because there are two bounds at work: the traditional ELBO (via Jensen's inequality) and the bound relating SDO to IB in Eq. (3.1.2). Our surrogate objective $\mathcal{L}_{\text{SDO}}(q)$ can also be seen as exactly the traditional ELBO applied directly to the IB model.

Our bound argument for applying the IB model to maximize the evidence of categorical data in Eq. (3.1.3) relies on crucial (but achievable) assumptions. First, both SDO and IB models need to have the *same* prior density π over weights (which should be Normal for conjugacy under the IB model). Second, that our approximate posterior Q with density q can conceptually generate the unknown variables for either SDO or IB. This is possible because given the same categorical data y, these distributions will have the same support.

Relation between DO and IB. The arguments above that justify the IB ELBO as a lower bound for the SDO ELBO would naturally hold for the original DO-Probit model as well, not just our simplified model, because the DO normalization constant is also guaranteed to be between 0 and 1. However, the posterior visualizations in Fig. 1 suggest

the SDO is a better target model at least for large N in the intercepts-only regime, as the IB's regions of high posterior density have more overlap with the SDO than the DO.

3.2 ALGORITHM FOR POSTERIOR ESTIMATION

Via the surrogate bound relation established in the previous section, we can provably optimize our intended categorical SDO model by instead maximizing the traditional ELBO of the independent binary model. This argument holds for any selected approximate posterior Q over $\boldsymbol{z}, \boldsymbol{\beta}$.

In this section, we develop an algorithm for coordinate ascent variational inference (CAVI) that can benefit from the conjugacy of the IB model. We assume that Q has a mean-field factorization: $q(\boldsymbol{z}, \boldsymbol{\beta}) = q(\boldsymbol{z})q(\boldsymbol{\beta})$. Under these choices of algorithm and factorization, deriving the steps of the algorithm follows a standard general recipe [Blei et al., 2017] and has been previously applied to binary probit models [Consonni and Marin, 2007, Armagan and Zaretzki, 2011, Fasano et al., 2021].

We can further generalize this strategy to *any* model $\mathcal{M}$ whose joint distribution over all visible and unobserved variables (denoted $\boldsymbol{u}$) includes an SDO-Probit complete data likelihood (Eq. (2.3.2)) for observed $\boldsymbol{y}$ and latent $\boldsymbol{z}$ as well as a conditionally-conjugate prior on weights $\boldsymbol{\beta}$. This generalization is possible since the lower bounding argument hinges solely on the expected complete data likelihood in Eq. (3.1.2). Generalization allows us to handle more flexible models such as the hierarchical SDO given in Section 4. We summarize our VI strategy in Algorithm 1.

Algorithm 1

1. Take $\mathcal{Q}$ to be a mean-field family with factorization: $q(u_1, ..., u_V) = \prod_{v=1}^{V} q_v(u_v)$, where each u_v is an unobserved variable ($\boldsymbol{z}$ and $\boldsymbol{\beta}$ are included in this set).
2. Construct the *surrogate model*, $\mathcal{M}_{\text{IB}}$, by swapping the IBP likelihood (2.2.3) for the SDO likelihood (2.3.2) and use it to compute *surrogate complete conditionals*: $\{\log p_{\mathcal{M}_{\text{IB}}}(u_v \mid \boldsymbol{u}_{-v}, \boldsymbol{y})\}$.
3. Define the objective: $\mathcal{L}_{\mathcal{M}}(q) = \mathbb{E}_q[\log \frac{p_{\mathcal{M}_{\text{IB}}}(\boldsymbol{y}, \boldsymbol{u})}{q(\boldsymbol{u})}]$.
4. Optimize $\mathcal{L}_{\mathcal{M}}(q)$ using optimal coordinate ascent updates [Blei et al., 2017]): $q_v(u_v) \propto \exp\left\{\mathbb{E}_{q_{-v}}\left[\log p_{\mathcal{M}_{\text{IB}}}(u_v \mid \boldsymbol{u}_{-v}, \boldsymbol{y})\right]\right\}$. If the complete conditional is an exponential family with natural parameter η_v, so is its optimal update, with natural parameter given by
$$\nu_v = \mathbb{E}_{q_{-v}}[\eta_v(\boldsymbol{u}_{-v}, \boldsymbol{y})] \tag{3.2.1}$$

The updates in Algorithm 1 will yield a density q^* that is a local maximum of the ELBO of the surrogate model [Ormerod and Wand, 2010], and therefore a local maximum of a surrogate bound on our intended SDO categorical model. All conditionals directly related to our SDO complete likelihood enjoy the closed-form updates given by Eq. (3.2.1). Note that the same strategy also applies to models with DO-Probit complete data likelihoods.

4 THE HIERARCHICAL SDO-PROBIT

We now illustrate the extensibility of our framework by introducing *hierarchical* SDO-Probit models, designed for datasets where observed categorical responses are nested within groups. For instance, the model can handle collections of categorical time series (see Section 4.1).

Let our dataset consist of J groups (indexed by j), each one containing N_j units (indexed by i). We assume that each group is given a *group-specific* weight parameter $\boldsymbol{\beta}_j$, and that these are generated in a hierarchical fashion with a common prior:

$$\begin{aligned}
\boldsymbol{\mu}_k &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{V}_0), && k = 1, ..., K \\
\boldsymbol{\Sigma}_k &\overset{\text{iid}}{\sim} \mathcal{W}^{-1}(\nu_0, \boldsymbol{S}_0), && k = 1, ..., K \\
\boldsymbol{\beta}_{jk} &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), && j = 1, ..., J
\end{aligned} \tag{4.0.1}$$

Here, we introduce unknown location vectors and covariance matrices $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ defining regression weights across groups for each category k, which have standard Normal and Inverse Wishart priors.

Given group level parameters $\boldsymbol{\beta}_j$, we can combine these with the covariates $\boldsymbol{x}_{ij}$ at each unit i within group j to produce categorical observations $y_{ij} \in \{1, 2, \dots K\}$:

$$y_{ij} \sim \text{Cat}(p_{ij1}, \dots p_{ijK}), \quad p_{ijk} = \frac{\Phi(\boldsymbol{x}_{ji}^T \boldsymbol{\beta}_{jk})}{\sum_{\ell=1}^{K} \Phi(\boldsymbol{x}_{ji}^T \boldsymbol{\beta}_{j\ell})}$$

Auxiliary variables $\boldsymbol{z}_{ij}$ can be also introduced from an appropriate truncated Normal, following the SDO generative process (Eq. (2.3.2)). Given this model, developing a coordinate ascent variational inference follows naturally from Algorithm 1. Details are worked out in Appendix C.

4.1 HIERARCHICAL AUTOREGRESSION

In order to model multiple related sequences of categorical observations over time, we develop a hierarchical multiclass autoregressive model as a special case of Eq. (4.0.1). In this case, the units (indexed by i) are timesteps, and the groups (indexed by j) are the sequences. We simply need to adjust the covariates so that they include the previous timestep's categorical outcome.

The dynamic covariates have the block structure

$$\boldsymbol{x}_{ji} = \left[(\boldsymbol{b}_{ji})^T \quad (\boldsymbol{e}_{y_{j,i-1}})^T \quad (\boldsymbol{x}_{ji}^{\text{exog}})^T\right]^T \in \mathbb{R}^M$$

The first block $\boldsymbol{b}_{ji}$ is $[1]$ if an intercept is used and empty otherwise. The second block $\boldsymbol{e}_{y_{j,i-1}}$ indicates the previous category $y_{j,i-1}$ via a one-hot vector of length K. The third block $\boldsymbol{x}^{\text{exog}} \in \mathbb{R}^{M_e}$ are exogenous dynamic covariates. We can write $M = 1 + K + M_e$, assuming an intercept is used.

Correspondingly, the regression weights decompose as

$$\boldsymbol{\beta}_{jk} := \left[(\boldsymbol{\beta}_{jk}^{\text{intercept}})^T \quad (\boldsymbol{\beta}_{jk}^{\text{transition}})^T \quad (\boldsymbol{\beta}_{jk}^{\text{exog}})^T\right]^T$$

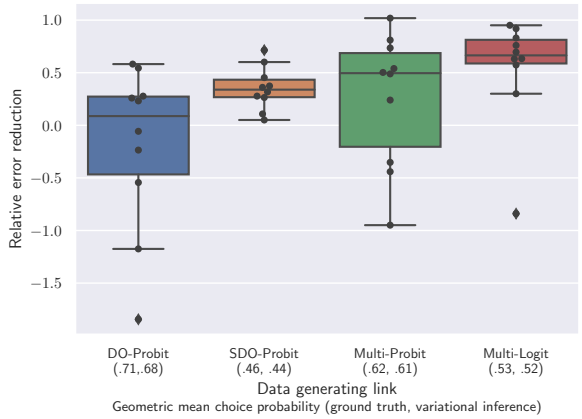


Figure 4: The reduction in relative error obtained by using the IB+SDO estimator over the IB+DO estimator after fitting the IB model with our variational inference scheme. Box plots show the spread in performance across several replicates of a data-generating process. Large positive values indicate that IB+SDO delivers better estimates.

Despite the complexities of this model, using our variational framework we can develop a coordinate ascent algorithm with *closed-form updates*.

5 EXPERIMENTS

5.1 PREDICTING *CATEGORICAL* RESPONSES WITH AN INDEPENDENT BINARY MODEL

Since the IB model via lower bound arguments approximates the posteriors of both the DO-Probit and SDO-Probit models, our goal is to assess in practice which model's estimator performs better at predicting heldout data.

The methodology for data simulation and model fitting is given in Sec. D.1. We consider synthetic datasets of $N = 4000$ training examples, $K = 3$ categories, and $M = 3$ covariates drawn from different ground truth models.[2] Given a variational posterior obtained via Algorithm 1, we compute the posterior mean weights as a point estimate, $\bar{\boldsymbol{\beta}}$, and then make predictions with either the IB+DO (Eq. (2.6.1)) or IB+SDO (Eq. (2.6.2)) estimators.

To assess the relative performance of estimators, we report the *relative reduction in error* from using IB+SDO over IB+DO as $(\bar{\ell}_{\text{SDO}} - \bar{\ell}_{\text{DO}})/(\bar{\ell}_{\text{DO}} - \bar{\ell}_{\text{GT}})$, where $\bar{\ell}$ is the mean log probability of the observed category across observations in the heldout test set. The ground truth likelihood, denoted GT, is obtained by plugging the known data-generating regression weights into the known model. We report the spread of these error reduction values across 10 replicates of each data-generating process in Fig. 4.

This figure suggests that IB+SDO produces superior performance on a typical dataset, although IB+DO is superior in certain cases. We thus recommend the IB+SDO estimator as a reasonable default. We assess absolute performance by reporting the geometric means $\left(\exp(\bar{\ell}_{\text{GT}}), \exp(\bar{\ell}_{\text{SDO}})\right)$ at the bottom of Figure 4, and find that our variational scheme provides a good approximation of the ground truth, despite intentionally using a mis-specified model.

[2]Our multinomial probit model used a diagonal covariance matrix on the latent variables. Thus, like the other models, it has the property of independence of the irrelevant alternative

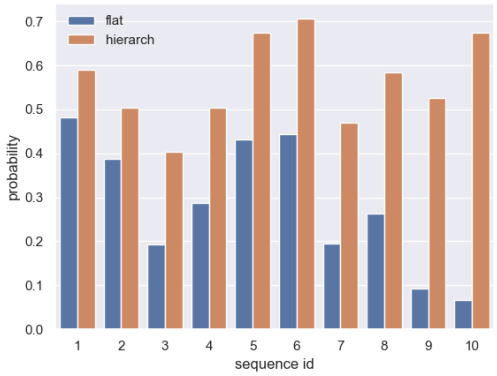


Figure 5: Mean test likelihood of the rarest category (5% of all observations). The hierarchical model assigns higher likelihood to this category when it appears.

5.2 MODELING MULTIPLE CATEGORICAL TIME SERIES

Here we demonstrate how applying Algorithm 1 to the autoregressive models of Sec. 4 enables easy variational inference for collections of categorical time series. We generated simulated data (as described in Sec. D.2) with 10 sequences, 5 categories, and 3 exogenous covariates. We fit a VI posterior approximating the hierarchical autoregressive SDO-Probit model of Sec. 4.1. Overall, we found that the hierarchical model provides a modest but consistent boost in heldout likelihood: the mean likelihood of test data increased from 59.9% to 62.5% with a minimum gain across groups of 1.3%. The hierarchical structure was *particularly* advantageous for predicting rare categories, as shown in Fig. 5. The hierarchical model shares statistical strength across time series, a capacity supported by our extensible framework.

6 DISCUSSION

In summary, our new variational methods provide a principled route to fast, scalable modeling of categorical data. While our variational methods are most practically viewed as simply fitting the conjugate binary model, we show this can be motivated via a rigorous lower bound on the marginal likelihood of a *categorical* model. Future work could apply these methods to real datasets, investigate the tightness of the objective, and extend these methods to models with latent (unobserved) categorical variables.

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. MCH is supported in part by NSF IIS-1908617.

References

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

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

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

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, and Jake VanderPlas. JAX: Composable transformations of Python+NumPy programs, 2018.

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

Lane F. Burgette, David Puelz, and P. Richard Hahn. A Symmetric Prior for Multinomial Probit Models. *Bayesian Analysis*, (-1):1–18, 2021.

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

Augusto Fasano, Daniele Durante, and Giacomo Zanella. Scalable and Accurate Variational Bayes for High-Dimensional Binary Regression Models. *arXiv:1911.06743 [stat]*, 2021.

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

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, page 31, 2014.

Tracy Holsclaw, Arthur M Greene, Andrew W Robertson, Padhraic Smyth, 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.

James Johndrow, David Dunson, and Kristian Lum. Diagonal orthant multinomial probit models. In *Artificial Intelligence and Statistics*, pages 29–38. PMLR, 2013.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

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

Måns Magnusson, Leif Jonsson, and Mattias Villani. DOLDA: A regularized supervised topic model for high-dimensional multi-class regression. *Computational Statistics*, 35(1):175–201, 2020.

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

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv:1912.11554 [cs, stat]*, 2019.

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

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

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

A PRELIMINARIES

A.1 THE INVERSE WISHART DISTRIBUTION

The Inverse Wishart is a distribution on symmetric, positive definite matrices. Under a natural parametrization, the Inverse Wishart distribution, denoted $\mathcal{W}^{-1}(\nu, \boldsymbol{\Psi})$, has density

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\right] \tag{A.1.1}$$

where $\boldsymbol{\Sigma} \succ 0$ and $\nu > d-1$ to have a proper prior. The expected value of an Inverse Wishart random variable parametrized as in (A.1.1) is given by $\mathbb{E}[\boldsymbol{\Sigma}] = \frac{\boldsymbol{\Psi}}{\nu-d-1}$. The expected value of the precision matrix is $\mathbb{E}[\boldsymbol{\Sigma}^{-1}] = \nu\boldsymbol{\Psi}^{-1}$.

A.2 UNIVARIATE NORMALS TRUNCATED TO POSITIVE OR NEGATIVE REALS

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 $T_- \sim N_-(\mu, 1)$. Based on this construction, it is straightforward to derive

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

$$\mathbb{E}[T_+] = \mu + \frac{\phi(-\mu)}{1-\Phi(-\mu)}, \quad \mathbb{E}[T_-] = \mu - \frac{\phi(-\mu)}{\Phi(-\mu)} \tag{A.2.1}$$

$$\text{Var}[T_+] = 1 - \mu(\mathbb{E}[T_+] - \mu) - (\mathbb{E}[T_+] - \mu)^2 \tag{A.2.2}$$

$$\text{Var}[T_-] = 1 - \mu(\mathbb{E}[T_-] - \mu) - (\mathbb{E}[T_-] - \mu)^2 \tag{A.2.3}$$

$$\mathbb{H}[T_+] = \ln\left(\sqrt{2\pi e}\,[1-\Phi(-\mu)]\right) - \frac{\mu\phi(-\mu)}{2(1-\Phi(-\mu))} \tag{A.2.4}$$

$$\mathbb{H}[T_-] = \ln\left(\sqrt{2\pi e}\,\Phi(-\mu)\right) + \frac{\mu\phi(-\mu)}{2\Phi(-\mu)} \tag{A.2.5}$$

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 A.2.1. (*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

$$\mathbb{E}[T_s] = \mu + \delta_s(\mu), \quad \delta_s(\mu) := \begin{cases} \dfrac{\phi(-\mu)}{1-\Phi(-\mu)}, & s = + \\ -\dfrac{\phi(-\mu)}{\Phi(-\mu)}, & s = - \end{cases} \tag{A.2.6}$$

which holds by (A.2.1). △

B ANOTHER MOTIVATION FOR THE IB-SDO ESTIMATOR

Another way to motivate the IB+SDO estimator is to consider an intercepts-only version of the IB model. Given a training set where the true frequency of category k is p_k, a simple "moment-matching" argument suggests that a good point estimate of $\bar{\boldsymbol{\beta}}$ should satisfy $\Phi(\bar{\boldsymbol{\beta}}_k) \approx p_k$. Thus, making predictions by computing $\Phi(\bar{\boldsymbol{\beta}}_k)$ for each category index k and normalizing should given a decent estimate of the training set's frequency of each category.

C VARIATIONAL INFERENCE FOR THE HIERARCHICAL SDO

C.1 VARIATIONAL FAMILY

We take the mean-field variational family for the Hierarchical SDO-probit model (4.0.1) to have density given by

$$q(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \overset{1}{=} q(\boldsymbol{\beta})q(\boldsymbol{z})q(\boldsymbol{\Sigma})q(\boldsymbol{\mu})$$

$$\overset{2}{=} \prod_{k=1}^{K} \underbrace{q(\boldsymbol{\mu}_k)}_{\mathcal{N}(\widetilde{\boldsymbol{m}}_k, \widetilde{\boldsymbol{V}}_k)} \underbrace{q(\boldsymbol{\Sigma}_k)}_{\mathcal{W}^{-1}(\widetilde{\nu}_k, \widetilde{\boldsymbol{S}}_k)} \prod_{j=1}^{J} \underbrace{q(\boldsymbol{\beta}_{jk})}_{\mathcal{N}(\widetilde{\boldsymbol{\mu}}_{jk}, \widetilde{\boldsymbol{\Sigma}}_{jk})} \prod_{i=1}^{N_j} \underbrace{q(z_{ijk})}_{\mathcal{TN}(\widetilde{\eta}_{ijk}, 1, \Omega_{ijk})} \tag{C.1.1}$$

$$\text{where } \Omega_{ijk} = \begin{cases} \mathbb{R}^+, & y_{ij} = k \\ \mathbb{R}^-, & \text{otherwise} \end{cases} \tag{C.1.2}$$

Equality (1) is by mean-field assumption, and, as we see below, (2) is the optimal such form as per Algorithm 1.

C.2 SURROGATE COMPLETE CONDITIONALS

The hierarchical Bayesian DO-Probit model (4.0.1) has a surrogate model which is a hierarchical Bayesian linear regression on the auxilliary variables $\boldsymbol{z}$. In this way, we obtain the surrogate complete conditionals

$$z_{ijk} \mid \boldsymbol{\beta}_{j1}, ..., \boldsymbol{\beta}_{jK}, y_{ij} \sim \begin{cases} \mathcal{N}_+\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{jk}\ ,\ 1\right), & y_{ij} = k \\ \mathcal{N}_-\left(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{jk}\ ,\ 1\right), & \text{otherwise} \end{cases} \tag{C.2.1}$$

where $\mathcal{N}_+$ and $\mathcal{N}_-$ are truncated normal distributions defined in Section A.2, and

$$\boldsymbol{\beta}_{jk} \mid \boldsymbol{z}_{jk}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \sim \mathcal{N}(\boldsymbol{\mu}'_{jk}, \boldsymbol{\Sigma}'_{jk}),$$

$$\boldsymbol{\mu}'_{jk} = \boldsymbol{\Sigma}'_{jk}\left(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{X}_j^T\boldsymbol{z}_{jk}\right), \quad \boldsymbol{\Sigma}'_{jk} = \left(\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{X}_j^T\boldsymbol{X}_j\right)^{-1} \tag{C.2.2}$$

and

$$\boldsymbol{\Sigma}_k \mid \boldsymbol{\mu}_k, \boldsymbol{\beta}_{1k}, ..., \boldsymbol{\beta}_{Jk} \sim \mathcal{W}^{-1}\left(\nu_0 + J, \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\mu}_k}\right),$$

$$\boldsymbol{S}_{\boldsymbol{\mu}_k} := \sum_{j=1}^{J}(\boldsymbol{\beta}_{jk} - \boldsymbol{\mu}_k)(\boldsymbol{\beta}_{jk} - \boldsymbol{\mu}_k)^T \tag{C.2.3}$$

and

$$\boldsymbol{\mu}_k \mid \boldsymbol{\Sigma}_k, \boldsymbol{\beta}_{1k}, ..., \boldsymbol{\beta}_{Jk} \sim \mathcal{N}(\boldsymbol{m}'_k, \boldsymbol{V}'_k),$$

$$\boldsymbol{m}'_k = \boldsymbol{V}'_k\left(\boldsymbol{V}_0\boldsymbol{m}_0 + J\boldsymbol{\Sigma}_k^{-1}\overline{\boldsymbol{\beta}_k}\right), \qquad \overline{\boldsymbol{\beta}_k} := \frac{1}{J}\sum_{j=1}^{J}\boldsymbol{\beta}_{jk}$$

$$\boldsymbol{V}'_k = \left(\boldsymbol{V}_0^{-1} + J\boldsymbol{\Sigma}_k^{-1}\right)^{-1} \tag{C.2.4}$$

C.3 COORDINATE ASCENT UPDATES

All of the surrogate complete conditionals are in the exponential family, and hence we know from Algorithm 1 that the optimal variational factors with respect to the surrogate lower bound $\mathcal{L}_{\text{H-SDO}}$ are in the same exponential family, with parameters given by (3.2.1). Here we derive the parameters for the updates, using the notation of (C.1.1).

Updates to $\{q(\boldsymbol{\beta}_{jk})\}_{jk}$ Since the natural parameters of a multivariate Gaussian are the precision and precision-weighted mean, we reparametrize the surrogate complete conditional for each $\boldsymbol{\beta}_{jk}$ in (C.2.2) before taking variational expectations of the parameters. Hence, the optimal update to each $q(\boldsymbol{\beta}_{jk} \mid \widetilde{\boldsymbol{\mu}}_{jk}, \widetilde{\boldsymbol{\Sigma}}_{jk})$ with respect to the objective $\mathcal{L}_{\text{H-SDO}}$ is given by

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_{jk}^{-1} &= \mathbb{E}_{q_{-\boldsymbol{\beta}_{jk}}}\left[\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{X}_j^T\boldsymbol{X}_j\right] = \mathbb{E}_{q_{\boldsymbol{\Sigma}_k}}[\boldsymbol{\Sigma}_k^{-1}] + \boldsymbol{X}_j^T\boldsymbol{X}_j \\
&\overset{1}{=} \widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\nu}_k + \boldsymbol{X}_j^T\boldsymbol{X}_j \\
\widetilde{\boldsymbol{\Sigma}}_{jk}^{-1}\widetilde{\boldsymbol{\mu}}_{jk} &= \mathbb{E}_{q_{-\boldsymbol{\beta}_{jk}}}\left[\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \boldsymbol{X}_j^T\boldsymbol{z}_{jk}\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\Sigma}_k}}[\boldsymbol{\Sigma}_k^{-1}]\mathbb{E}_{q_{\boldsymbol{\mu}_k}}[\boldsymbol{\mu}_k] + \boldsymbol{X}_j^T\mathbb{E}_{q_{\boldsymbol{z}_{jk}}}[\boldsymbol{z}_{jk}] \\
&\overset{2}{=} \widetilde{\nu}_k\widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\boldsymbol{m}}_k + \boldsymbol{X}_j^T\mathbb{E}_{q_{\boldsymbol{z}_{jk}}}[\boldsymbol{z}_{jk}]
\end{aligned}$$

where $\mathbb{E}_q[\boldsymbol{z}_{jk}]$ is given explicitly below. Equations (1), (2) follow from Sec. A.1.

Thus, in standard parameterization, we update

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_{jk} &= \widetilde{\boldsymbol{\Sigma}}_{jk}\left(\widetilde{\nu}_k\widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\boldsymbol{m}}_k + \boldsymbol{X}_j^T\mathbb{E}_{q_{\boldsymbol{z}_{jk}}}[\boldsymbol{z}_{jk}]\right) \\
\widetilde{\boldsymbol{\Sigma}}_{jk} &= \left(\widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\nu}_k + \boldsymbol{X}_j^T\boldsymbol{X}_j\right)^{-1}
\end{aligned}$$

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

$$\mathbb{E}_q[z_{ijk}] = \begin{cases} \widetilde{\eta}_{ijk} + \dfrac{\phi(-\widetilde{\eta}_{ijk})}{1-\Phi(-\widetilde{\eta}_{ijk})}, & y_{ij} = k \\ \widetilde{\eta}_{ijk} - \dfrac{\phi(-\widetilde{\eta}_{ijk})}{\Phi(-\widetilde{\eta}_{ijk})}, & \text{otherwise} \end{cases}$$

by properties of the truncated normal distribution (Section A.2). Recall that ϕ and Φ refer to the pdf and cdf, respectively, of the standard normal.

Updates to $\{q(\boldsymbol{\mu}_k)\}_k$ Since the natural parameters of a multivariate Gaussian are the precision and precision-weighted mean, we reparametrize the surrogate complete conditionals for $\boldsymbol{\mu}_k$ in (C.2.4) before taking variational expectations of the parameters. Hence, the optimal update to $q(\boldsymbol{\mu}_k \mid \widetilde{\boldsymbol{m}}_k, \widetilde{\boldsymbol{V}}_k)$ with respect to the objective $\mathcal{L}_{\text{H-SDO}}$ is given by

$$\begin{aligned}
\widetilde{\boldsymbol{V}}_k^{-1} &= \mathbb{E}_{q_{-\boldsymbol{\mu}+k}}\left[\boldsymbol{V}_0^{-1} + J\boldsymbol{\Sigma}_k^{-1}\right] = \boldsymbol{V}_0^{-1} + J\,\mathbb{E}_{q_{\boldsymbol{\Sigma}_k}}[\boldsymbol{\Sigma}_k^{-1}] \\
&\overset{1}{=} \boldsymbol{V}_0^{-1} + J\widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\nu}_k \\
\widetilde{\boldsymbol{V}}_k^{-1}\widetilde{\boldsymbol{m}}_k &= \mathbb{E}_{q_{-\boldsymbol{\mu}_k}}\left[\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + J\boldsymbol{\Sigma}_k^{-1}\sum_{j=1}^J\boldsymbol{\beta}_{jk}\right] \\
&= \boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + \cancel{J}\,\mathbb{E}_{q_{\boldsymbol{\Sigma}_k}}[\boldsymbol{\Sigma}_k^{-1}]\,\frac{1}{\cancel{J}}\sum_{j=1}^J\mathbb{E}_{q_{\boldsymbol{\beta}_{jk}}}[\boldsymbol{\beta}_{jk}] \\
&\overset{2}{=} \boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + \widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\nu}_k\sum_{j=1}^J\widetilde{\boldsymbol{\mu}}_{jk}
\end{aligned}$$

Equations (1), (2) follow from Sec. A.1.

Thus, in standard parameterization, we update

$$\widetilde{\boldsymbol{m}}_k = \widetilde{\boldsymbol{V}}_k\left(\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + \widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\nu}_k\sum_{j=1}^J\widetilde{\boldsymbol{\mu}}_{jk}\right), \quad \widetilde{\boldsymbol{V}}_k = \left(\boldsymbol{V}_0^{-1} + J\widetilde{\boldsymbol{S}}_k^{-1}\widetilde{\nu}_k\right)^{-1}$$

Updates to $\{q(\boldsymbol{\Sigma}_k)\}_k$ Since the natural parameters of an Inverse Wishart are identical to the conventional parameters (up to a multiplicative scalar constant), we do not need to reparametrize the surrogate complete conditional for $\boldsymbol{\Sigma}_k$ in (C.2.3) before taking variational expectations of the parameters. Hence, the optimal update to $q(\boldsymbol{\Sigma}_k \mid \widetilde{\boldsymbol{S}}_k, \widetilde{\nu}_k)$ with respect to the objective $\mathcal{L}_{\text{H-SDO}}$ is given by

$$\begin{aligned}
\widetilde{\nu}_k &= \mathbb{E}_{q_{-\boldsymbol{\Sigma}}}\left[\nu_0 + J\right] = \nu_0 + J \\
\widetilde{\boldsymbol{S}}_k &= \mathbb{E}_{q_{-\boldsymbol{\Sigma}}}\left[\boldsymbol{S}_0 + \sum_{j=1}^J(\boldsymbol{\beta}_{jk}-\boldsymbol{\mu}_k)(\boldsymbol{\beta}_{jk}-\boldsymbol{\mu}_k)^T\right] \\
&= \boldsymbol{S}_0 + \sum_{j=1}^J\mathbb{E}_{q_{-\boldsymbol{\Sigma}}}\left[\boldsymbol{\beta}_{jk}\boldsymbol{\beta}_{jk}^T - \boldsymbol{\beta}_{jk}\boldsymbol{\mu}_k^T + \boldsymbol{\mu}_k\boldsymbol{\beta}_{jk}^T + \boldsymbol{\mu}_k\boldsymbol{\mu}_k^T\right] \\
&\overset{(+)}{=} \boldsymbol{S}_0 + J\widetilde{\boldsymbol{V}}_k + \sum_{j=1}^J\widetilde{\boldsymbol{\Sigma}}_{jk} + (\widetilde{\boldsymbol{\mu}}_{jk}-\widetilde{\boldsymbol{m}}_k)(\widetilde{\boldsymbol{\mu}}_{jk}-\widetilde{\boldsymbol{m}}_k)^T
\end{aligned} \tag{C.3.1}$$

where in (+) we used

$$\mathbb{E}_q[\boldsymbol{\mu}_k\boldsymbol{\mu}_k^T] = \text{Var}_q[\boldsymbol{\mu}_k] + \mathbb{E}_q[\boldsymbol{\mu}_k]\mathbb{E}_q[\boldsymbol{\mu}_k]^T = \widetilde{\boldsymbol{V}}_k + \widetilde{\boldsymbol{m}}_k\widetilde{\boldsymbol{m}}_k^T$$
$$\mathbb{E}_q[\boldsymbol{\beta}_{jk}\boldsymbol{\beta}_{jk}^T] = \text{Var}_q[\boldsymbol{\beta}_{jk}] + \mathbb{E}_q[\boldsymbol{\beta}_{jk}]\mathbb{E}_q[\boldsymbol{\beta}_{jk}]^T = \widetilde{\boldsymbol{\Sigma}}_{jk} + \widetilde{\boldsymbol{\mu}}_{jk}\widetilde{\boldsymbol{\mu}}_{jk}^T$$

and

$$\mathbb{E}_q[\boldsymbol{\beta}_{jk}\boldsymbol{\mu}_k^T] \overset{\text{(mean field)}}{=} \mathbb{E}_q[\boldsymbol{\beta}_{jk}]\mathbb{E}_q[\boldsymbol{\mu}_k^T] = \widetilde{\boldsymbol{\mu}}_{jk}\widetilde{\boldsymbol{m}}_k^T$$

with $\mathbb{E}_q[\boldsymbol{\mu}_k\boldsymbol{\beta}_{jk}^T]$ handled similarly as $\mathbb{E}_q[\boldsymbol{\beta}_{jk}\boldsymbol{\mu}_k^T]$.

Updates to $\{q(z_{ijk})\}_{ijk}$ In (C.2.1) , we saw that the surrogate complete conditional for each z_{ijk} has the form $\mathcal{TN}(\eta_{ijk}, 1, \Omega_{ijk})$, where Ω_{ijk} is defined as in (C.1.2). But since each such distribution is in the exponential family with natural parameter η_{ijk}, the optimal update for each $q(z_{ijk})$ is given by

$$\widetilde{\boldsymbol{\eta}}_{ijk} = \mathbb{E}[\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_{jk}] = \boldsymbol{x}_{ij}^T\widetilde{\boldsymbol{\mu}}_{jk}$$

C.4 LOWER BOUND

A tractable surrogate lower bound on the marginal log likelihood for the Hierarchical DO-Probit model (4.0.1) – useful for tracking the monotonic increase in the objective function and setting a convergence criterion – can immediately be obtained by handling the expected complete data likelihood precisely as in Section 3.1

$$\begin{aligned}
&\texttt{ELBO}_{\text{H-SDO}}(q; \boldsymbol{y}) = \underbrace{\mathbb{E}_q[\log p_{\text{H-SDO}}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})]}_{\text{energy}} \underbrace{-\mathbb{E}_q[\log q(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})]}_{\text{entropy}} \\
&= \mathbb{E}_q[\log p_{\text{SDO}}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta})] + \mathbb{H}[q(\boldsymbol{z})] - \texttt{KL}\big(q(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,||\, p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\big) \\
&\overset{(3.1.2)}{\geq} \mathbb{E}_q[\log p_{\text{IB}}(\bar{\boldsymbol{y}}, \boldsymbol{z} \mid \boldsymbol{\beta})] + \mathbb{H}[q(\boldsymbol{z})] - \texttt{KL}\big(q(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \,||\, p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\big) \\
&= \texttt{ELBO}_{\text{IB}}(q, \bar{\boldsymbol{y}}) - \texttt{KL}\big(q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \,||\, p(\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) := \mathcal{L}_{\text{H-SDO}}(q; \boldsymbol{y}) \qquad \text{(C.4.1)}
\end{aligned}$$

We expand this lower bound using:

$$\begin{aligned}
\mathcal{L}_{\text{H-SDO}}(q) = &\sum_{j=1}^{J}\sum_{i=1}^{N_j}\sum_{k=1}^{K} \underbrace{\mathbb{E}_q[\log p_{\text{IB}}(\bar{y}_{ijk}, z_{ijk} \mid \boldsymbol{\beta}_{jk})]}_{\text{(A)}} \\
&+ \sum_{j=1}^{J}\sum_{i=1}^{N_j}\sum_{k=1}^{K} \underbrace{\mathbb{H}[q(z_{ijk})]}_{\text{(B)}} + \sum_{k=1}^{K} \underbrace{-\texttt{KL}\big(q(\boldsymbol{\mu}_k) \,||\, p_{\text{H-SDO}}(\boldsymbol{\mu}_k)\big)}_{\text{(C)}} \\
&+ \sum_{k=1}^{K} \underbrace{-\texttt{KL}\big(q(\boldsymbol{\Sigma}_k) \,||\, p_{\text{H-SDO}}(\boldsymbol{\Sigma}_k)\big)}_{\text{(D)}} \\
&+ \sum_{j=1}^{J}\sum_{k=1}^{K} \underbrace{-\mathbb{E}_q\Big[\texttt{KL}\big(q(\boldsymbol{\beta}_{jk}) \,||\, p_{\text{H-SDO}}(\boldsymbol{\beta}_{jk} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\big)\Big]}_{\text{(E)}}.
\end{aligned}$$

In particular term (A) is given by

$$\begin{aligned}
&\mathbb{E}_q[\log p_{\text{IB}}(\bar{y}_{ijk}, z_{ijk} \mid \boldsymbol{\beta}_{jk})] \\
&\quad = -\frac{1}{2}(\log 2\pi + 1) + \frac{1}{2}\widetilde{\eta}_{ijk}\delta_{\bar{y}_{ijk}}(\widetilde{\eta}_{ijk}) - \frac{1}{2}\boldsymbol{x}_{ij}^T\widetilde{\boldsymbol{\Sigma}}_k\boldsymbol{x}_{ij}
\end{aligned}$$

where

$$\widetilde{\eta}_{ijk} = \boldsymbol{x}_{ij}^T\widetilde{\boldsymbol{\mu}}_k \quad \text{and} \quad \delta_b(x) := \begin{cases} \dfrac{\phi(-x)}{1-\Phi(-x)}, & b = 1 \\ -\dfrac{\phi(-x)}{\Phi(-x)}, & b = 0 \end{cases}$$

Term (B) is the entropy of a truncated normal distribution, where the (ijk)-th element in the sum is

- the entropy of $\mathcal{N}_+(\boldsymbol{x}_{ij}^T\widetilde{\boldsymbol{\mu}}_k, 1)$ when $y_{ij} = k$, in which case the entropy is given by (A.2.4)
- the entropy of $\mathcal{N}_-(\boldsymbol{x}_{ij}^T\widetilde{\boldsymbol{\mu}}_k, 1)$ when $y_{ij} \neq k$, in which case the entropy is given by (A.2.5).

Term C is the KL divergence between two multivariate Gaussians; in this case we have

$$\begin{aligned}
&\texttt{KL}(q(\boldsymbol{\mu}_k) \,||\, p(\boldsymbol{\mu}_k)) \\
&= -\frac{1}{2}\Big[\log\frac{|\boldsymbol{V}_0|}{|\widetilde{\boldsymbol{V}}_k|} - M + (\widetilde{\boldsymbol{m}}_k - \boldsymbol{m}_0)^T\boldsymbol{V}_0^{-1}(\widetilde{\boldsymbol{m}}_k - \boldsymbol{m}_0) + \text{tr}(\boldsymbol{V}_0^{-1}\widetilde{\boldsymbol{V}}_k)\Big]
\end{aligned}$$

Term D is the KL divergence between two Inverse Wisharts; in this case, we have

$$\begin{aligned}
&\texttt{KL}\big(q(\boldsymbol{\Sigma}_k) \,||\, p(\boldsymbol{\Sigma}_k)\big) = \log\left(\frac{\Gamma_M(\frac{\nu_0}{2})}{\Gamma_M(\frac{\widetilde{\nu}_k}{2})}\right) + \frac{\widetilde{\nu}_k}{2}\text{tr}(\widetilde{\boldsymbol{S}}_k^{-1}\boldsymbol{S}_0) \\
&\quad - \frac{\widetilde{\nu}_k M}{2} - \frac{\nu_0}{2}\log|\widetilde{\boldsymbol{S}}_k^{-1}\boldsymbol{S}_0| - \frac{\nu_0 - \widetilde{\nu}_k}{2}\sum_{i=1}^{M}\psi\left(\frac{\widetilde{\nu}_k - M + i}{2}\right)
\end{aligned}$$

where Γ_M is the multivariate gamma function and ψ is the digamma function given by $\psi(x) := \frac{d}{dx}\ln\Gamma_1(x)$.

Term E gives the expected KL divergence between two multivariate Gaussians, in the specific case where the expectation is taken with respect to independent Gaussian and Inverse Wishart distributions on the parameters of the second argument of the KL divergence. Utilizing the mean-field assumption, we find

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\mu}_k \mid \widetilde{\boldsymbol{m}}_k, \widetilde{\boldsymbol{V}}_k)q(\boldsymbol{\Sigma}_k \mid \widetilde{\nu}_k, \widetilde{\boldsymbol{S}}_k)}\Big[\texttt{KL}\big(q(\boldsymbol{\beta}_{jk} \mid \widetilde{\boldsymbol{\mu}}_{jk}, \widetilde{\boldsymbol{\Sigma}}_{jk}) \,||\, p(\boldsymbol{\beta}_{jk} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\big)\Big] \\
&= \frac{1}{2}\Bigg[-M(\log 2 + 1) + \log|\widetilde{\boldsymbol{S}}_k| - \sum_{i=1}^{M}\psi\left(\frac{\widetilde{\nu}_k - i + 1}{2}\right) - \log|\widetilde{\boldsymbol{\Sigma}}_{jk}| \\
&\quad + \text{tr}\Big(\widetilde{\nu}_k\widetilde{\boldsymbol{\Psi}}_k^{-1}\big[\widetilde{\boldsymbol{\mu}}_{jk}\widetilde{\boldsymbol{\mu}}_{jk}^T - \widetilde{\boldsymbol{\mu}}_{jk}\widetilde{\boldsymbol{m}}_k^T - \widetilde{\boldsymbol{m}}_k\widetilde{\boldsymbol{\mu}}_{jk}^T + \widetilde{\boldsymbol{V}}_k + \widetilde{\boldsymbol{m}}_k\widetilde{\boldsymbol{m}}_k^T + \widetilde{\boldsymbol{\Sigma}}_{jk}\big]\Big)\Bigg].
\end{aligned}$$

D METHODOLOGY FOR EXPERIMENTS

All models in both experiments were run to a convergence criterion satisfied when consecutive iterations produced a drop of 1.0 or less in the surrogate lower bound on the log-evidence.

D.1 METHODOLOGY FOR EXPERIMENT 1

We generated a toy dataset of iid categorical observations as follows. We randomly sampled regression weights $\boldsymbol{\beta}_k \overset{\text{iid}}{\sim} \mathcal{N}_M(\boldsymbol{0}, \boldsymbol{I})$ for each of $k = 1, ..., K$ categories ($K = 3$) and $M = 4$ covariates. For each of $N = 5,000$ samples we sampled exogenous covariates $\boldsymbol{x}_i^{\text{exog}} \sim \mathcal{N}_3(\boldsymbol{0}, \boldsymbol{I})$ and added an intercept to obtain $\boldsymbol{x}_i = (1, \boldsymbol{x}_i^{\text{exog}})$. Given the covariates and regression weights, we then sampled categorical observations $y_i \in \{1, ..., K\}$ according to three different generative models: the DO-Probit, SDO-Probit, and Multinomial Logit model.

We performed variational inference using Algorithm 1 for the SDO-Probit model. We trained on the first 4,000 observations and tested on the remaining 1,000.

D.2 METHODOLOGY FOR EXPERIMENT 2

We generated a toy collection of categorical sequences as follows. We assumed $K = 5$ categorical responses. For covariates, we assumed $B = 1$ intercept term, K (first-order) autoregessive terms, and $M_e = 3$ exogenous covariates; thus, each $\boldsymbol{x}_{ji}$ had $M = B + K + M_e = 9$ elements. We assumed no interactions.

We generated data for $J = 10$ sequences ($T_j^{\text{train}} \equiv 50$ training timesteps and $T_j^{\text{test}} \equiv 950$ testing timesteps) via the generating process

$$
\begin{aligned}
\boldsymbol{\mu} &= \begin{bmatrix}
-2. & 0. & 0. & 0. & 0. \\
4. & 0. & 0. & 0. & 0. \\
-0.348 & -0.238 & 2.184 & -0.251 & -0.12 \\
1.411 & -0.254 & 0.969 & 0.936 & -1.34 \\
0. & 0. & 0. & 0. & 0. \\
0. & 0.1 & 0. & 0. & 0. \\
0. & 0. & 0.1 & 0. & 0. \\
0. & 0. & 0. & 0.1 & 0. \\
0. & 0. & 0. & 0. & 0.1
\end{bmatrix} \\
\boldsymbol{\Sigma}_k &\equiv \sigma^2 \boldsymbol{I}_{M \times M}, \quad \sigma^2 = 1.0 \\
\boldsymbol{\beta}_{jk} &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
\boldsymbol{x}_{ji} &= \begin{bmatrix} 1 & x_{ji1} & x_{ji2} & x_{ji3} & \boldsymbol{e}_{y_{j,i-1}} \end{bmatrix} \\
x_{ji1} &\overset{\text{iid}}{\sim} \text{Bernoulli}(p = .05) \\
x_{jim} &\overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \quad m \in \{2, 3\} \\
y_{ij} &= k \mid \boldsymbol{\beta}_j \;\; \text{according to (2.1.4)}
\end{aligned}
$$

where $\boldsymbol{\mu}_k$ is the kth column of the matrix $\boldsymbol{\mu} \in \mathbb{R}^{M \times K}$. Note that $\boldsymbol{\mu}$ discourages the appearance of the 1st category, but that the rare binary covariate x_{ji1} is highly predictive of its appearance.

We performed inference with a hierarchical DO-Probit autoregression (4.0.1) (see Section 4.1), as well as a collection of separate flat DO-Probit autoregressions (2.1.1), one for each group.

After training, we computed the likelihood (i.e. model probabilities) of categorical observations in a hold-out test set. Category probabilities were computed using (2.1.4) with $\mathbb{E}_q[\boldsymbol{\beta} \mid \boldsymbol{y}]$.