Easy Variational Inference for Categorical Models via an Independent Binary Approximation

Michael T. Wojnowicz, Shuchin Aeron, Eric L. Miller, and Michael C. Hughes

July 20, 2022

Context

Bayesian Categorical Generalized Linear Models (GLMs)

Given N observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ of covariates $\boldsymbol{x}_i$ and categorical outcomes $y_i \in \{1, \ldots, K\}$, we assume

$$\underbrace{\boldsymbol{p}_i}_{\text{probability vector}} = f\left(\underbrace{\boldsymbol{B}^T}_{\text{regression weights}} \underbrace{\boldsymbol{x}_i}_{\text{covariates}} \right) \quad \text{where } f \text{ (e.g. softmax) maps } \mathbb{R}^K \text{ to the simplex}$$

$$\underbrace{y_i}_{\text{outcome}} \sim \text{Categorical}_K\Big(\underbrace{\boldsymbol{p}_i}_{\text{probability vector}} \Big)$$

Goal: Given a prior on $\boldsymbol{B}$, we want to estimate the posterior.

Context

Bayesian Categorical Generalized Linear Models (GLMs)

Given N observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ of covariates $\boldsymbol{x}_i$ and categorical outcomes $y_i \in \{1, \ldots, K\}$, we assume

$$\underbrace{\boldsymbol{p}_i}_{\text{probability vector}} = f\left(\underbrace{\boldsymbol{B}^T}_{\text{regression weights}} \underbrace{\boldsymbol{x}_i}_{\text{covariates}} \right) \quad \text{where } f \text{ (e.g. softmax) maps } \mathbb{R}^K \text{ to the simplex}$$

$$\underbrace{y_i}_{\text{outcome}} \sim \text{Categorical}_K\Big(\underbrace{\boldsymbol{p}_i}_{\text{probability vector}} \Big)$$

Goal: Given a prior on $\boldsymbol{B}$, we want to estimate the posterior.

1. Different choices of f give different categorical GLMs with different properties for Bayesian inference.
2. We will define a family of f that has nice properties.

Bayesian inference with existing models

Table 1: Assessment of categorical GLMs in terms of the presence (✓) or absence (✗) of desirable features for simple, fast, scalable Bayesian inference

Model	**Inference Feature**				
	Conditional conjugacy	Closed-form variational inference	Invariance to category ordering	Closed-form category probabilities	Embarassingly parallel across categories
Softmax	✗	✗	✓	✓	✗
Multinomial Probit	✗	✗	✓	✗	✗
Softmax + Pòlya-Gamma aug.	✓	✗	✓	✓	✗
Stickbreaking-Softmax + Pòlya-Gamma aug.	✓	✓	✗	✓	✗
Multinomial Probit + Albert-Chib aug.	✓	✓	✓	✗	✗

Bayesian inference with existing models

Table 1: Assessment of categorical GLMs in terms of the presence (✓) or absence (✗) of desirable features for simple, fast, scalable Bayesian inference

Model	**Inference Feature**				
	Conditional conjugacy	Closed-form variational inference	Invariance to category ordering	Closed-form category probabilities	Embarassingly parallel across categories
Softmax	✗	✗	✓	✓	✗
Multinomial Probit	✗	✗	✓	✗	✗
Softmax + Pòlya-Gamma aug.	✓	✗	✓	✓	✗
Stickbreaking-Softmax + Pòlya-Gamma aug.	✓	✓	✗	✓	✗
Multinomial Probit + Albert-Chib aug.	✓	✓	✓	✗	✗

We pursue Bayesian inference for categorical GLMs that has all 5 features (and more).

Our approach: A new class of categorical GLMs

Independent binary (IB) models **[a model giving easy Bayesian inference]**

Let each observation $\widehat{\boldsymbol{y}}_i = (\widehat{y}_{i1}, ..., \widehat{y}_{iK})$ be a K-bit vector. (Example: K=3, $\widehat{\boldsymbol{y}}_i = (1, 0, 1)$.) The IB likelihood is the product of K binary regression likelihoods

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

where the k-th bit has a success probability $\rho_{ik} = H(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k)$ obtained by applying any univariate cdf H (e.g., Logistic, Gaussian, etc.) to the dot product of covariates $\boldsymbol{x}_i$ and regression weights $\widehat{\boldsymbol{\beta}}_k$ specific to the k-th bit.

Our approach: A new class of categorical GLMs

Independent binary (IB) models [a model giving easy Bayesian inference]

Let each observation $\widehat{\boldsymbol{y}}_i = (\widehat{y}_{i1}, ..., \widehat{y}_{iK})$ be a K-bit vector. (Example: K=3, $\widehat{\boldsymbol{y}}_i = (1, 0, 1)$.) The IB likelihood is the product of K binary regression likelihoods

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

where the k-th bit has a success probability $\rho_{ik} = H(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k)$ obtained by applying any univariate cdf H (e.g., Logistic, Gaussian, etc.) to the dot product of covariates $\boldsymbol{x}_i$ and regression weights $\widehat{\boldsymbol{\beta}}_k$ specific to the k-th bit.

Categorical-from-binary (CB) models [a new class of categorical GLMs]

CB models are GLMs for categorical data which obey the likelihood bound

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

where $\boldsymbol{e}_{y_i}$ is the one-hot indicator vector with value of 1 only at entry y_i.

Inference

- Closed-form coordinate ascent variational inference (CAVI) already exists for binary models and is easy, fast, and scalable.[1]
- So closed-form CAVI on IB models – which we call IB-CAVI (Independent Binary Coordinate Ascent Variational Inference) – exists with the same properties.
- By a quick argument, we see that IB-CAVI can be viewed as maximizing the marginal likelihood of our CB models.

[1] Consonni and Marin, (2007), *Computational Statistics & Data Analysis*; Durante and Rigon (2019), *Statistical Science.*

Inference

Opportunity: The posterior from IB-CAVI estimates the posterior of multiple CB models. To illustrate, suppose there are $K = 3$ categories.

1. *Categorical-from-binary-via-marginalization* (CBM) models normalize the marginal probabilities of success.

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

2. *Categorical-from-binary-via-conditioning* (CBC) models condition on the event that the K-bit vector has exactly one positive entry.

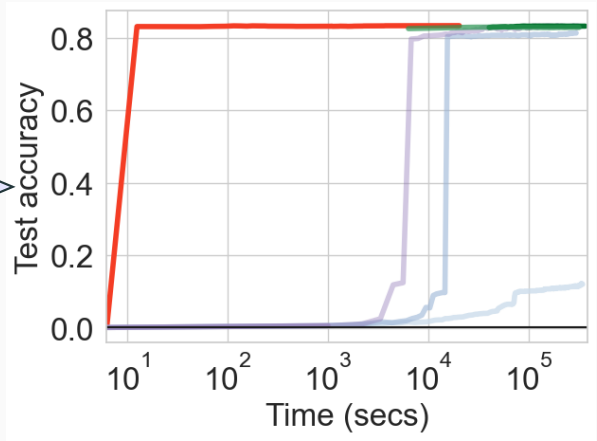
$$p_{\text{CBC}}(y = 1 \mid \boldsymbol{B}) = p_{\text{IB}}\left(\widehat{\boldsymbol{y}} = (1,0,0) \,\middle|\, \widehat{\boldsymbol{y}} \in \left\{(1,0,0),(0,1,0),(0,0,1)\right\}, \boldsymbol{B}\right)$$

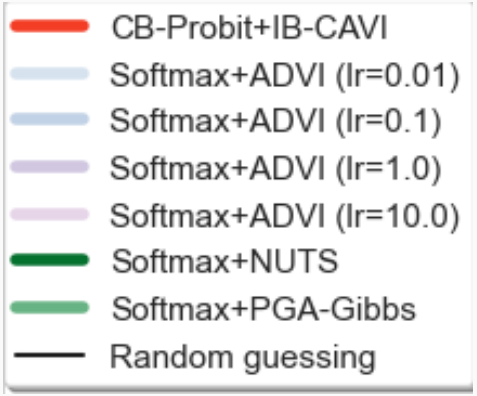
Bayesian model averaging improves inference quality!

Evaluation

We find that IB-CAVI (with BMA) delivers *similar predictive performance* as alternatives, while requiring *far less time* to get there.

Indistinguishable accuracy.





Little-to-no cost in log likelihood.

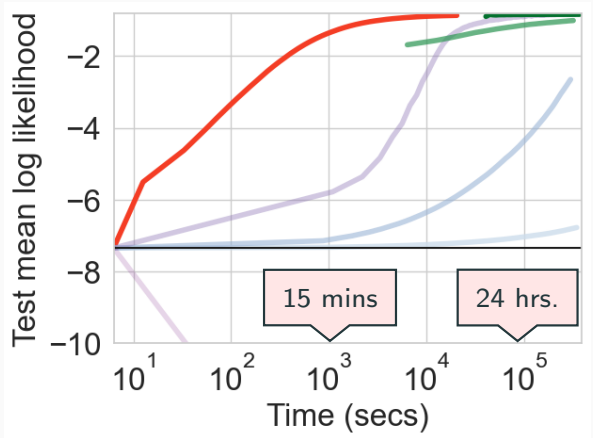
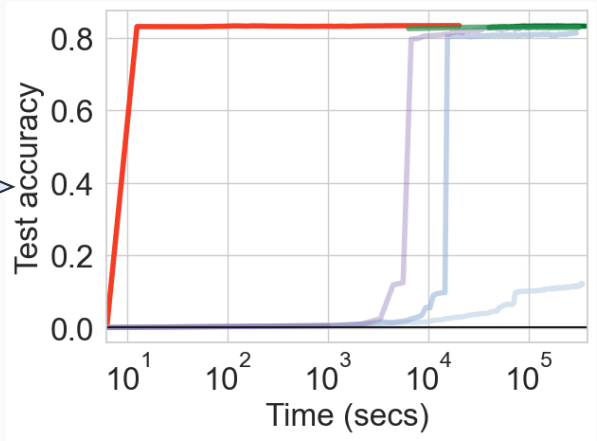


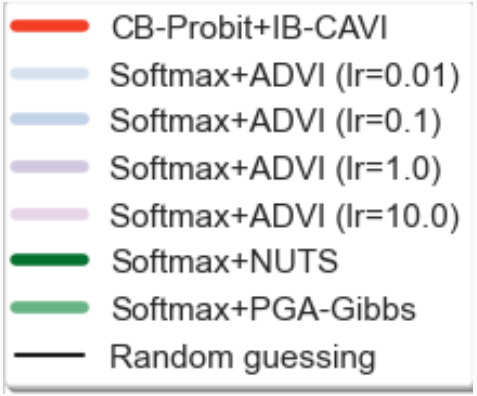
Figure 1: Predicting a computer user's process starts (with 1,553 categories, 1,553 covariates, and 17,724 examples) in an intruder detection experiment.

Evaluation

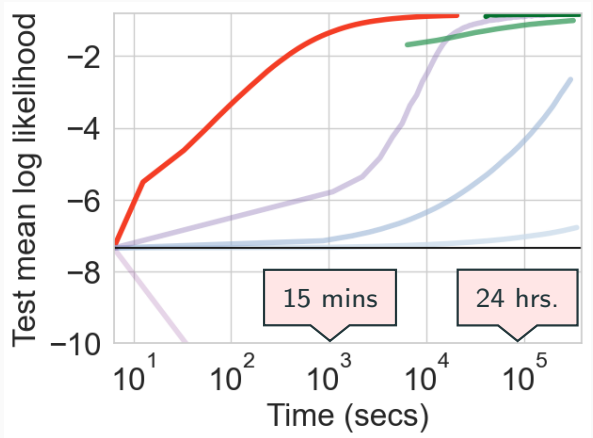
We find that IB-CAVI (with BMA) delivers *similar predictive performance* as alternatives, while requiring *far less time* to get there.

Indistinguishable accuracy.





Little-to-no cost in log likelihood.



These are the results even *before* exploiting parallelism.

Figure 1: Predicting a computer user's process starts (with 1,553 categories, 1,553 covariates, and 17,724 examples) in an intruder detection experiment.