5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–13

Approximate inference by broadening the support of the likelihood

Anonymous Authors
Anonymous Institution

Abstract

Here we present a framework for approximate inference on a target observation model F via inference on an observation model H with broader support. In many scenarios, there is a model H with broadened support which gives relatively easy and efficient inference. For example, inference is often easier with an independent binary model than with a categorical model, or with an unconstrained model than with a model truncated to some region. If the pair (F, H) is chosen such that the likelihood of F dominates that of H, then our framework gives a recipe for approximate inference. In the frequentist paradigm, we can substitute a parameter learned with H by maximum likelihood into F. In the Bayesian paradigm, we can treat the posterior using H as an approximate posterior using F. Our framework also provides the asymptotic discrepancy induced by the approximation.

1. Introduction

Suppose we observe the random variables $Y_i \overset{\text{ind}}{\sim} G$ for $i = 1, \ldots, n$, where G is an unknown probability distribution on $\mathcal{Y}$. Suppose further that we wish to model our observations as $Y_i \overset{\text{ind}}{\sim} F_\theta$, where $\{F_\theta : \theta \in \Theta\}$ is a family of probability distributions on $\mathcal{Y}$ indexed by parameter θ, and where each F_θ has density f_θ w.r.t. some σ-finite measure μ on the measurable space $(\mathcal{Y}, \mathcal{F})$.[1,2] When inference with the desired model F_θ is intractable (often due to intractable normalizing constants; see Sec. 3 for examples), it can be convenient to do approximate inference by artificially broadening the support of the likelihoods. In particular, at inference time, we assume that $Y_i \overset{\text{ind}}{\sim} H_\phi$, where $\{H_\phi : \phi \in \Phi\}$ is a family of probability distributions indexed by parameter ϕ on a space $(\mathcal{Y}*, \mathcal{F}*)$ where $\mathcal{Y}* \supsetneq \mathcal{Y}, \mathcal{F}* \supsetneq \mathcal{F}$, where each H_ϕ has density h_ϕ w.r.t. some σ-finite measure ν, and where

$$\texttt{supp}(H_\phi) \supsetneq \texttt{supp}(F_\theta), \qquad \forall \phi \in \Phi, \theta \in \Theta. \qquad \textit{(Assumption 1)} \tag{1.0.1}$$

In a typical application, we hope that the observation model is well-specified (i.e. that $F_\theta = G$ for some θ), but we recognize that this is unlikely. However, we can typically have more confidence in the more innocuous assumption of *well-specified support*

$$\texttt{supp}(F_\theta) = \texttt{supp}(G), \qquad \forall \theta \in \Theta \qquad \textit{(Assumption 2)} \tag{1.0.2}$$

[1] In particular if μ is Lebesgue measure, then f becomes a probability density function (pdf) of an absolutely continuous random variable. If μ is the counting measure, then f becomes a probability mass function (pmf) of a discrete random variable.

[2] To ease notation, we write $Y_i \overset{\text{ind}}{\sim} G$ instead of $Y_i \overset{\text{ind}}{\sim} G_i$, $Y_i \overset{\text{ind}}{\sim} F_\theta$ instead of $Y_i \overset{\text{ind}}{\sim} F_{i,\theta}$, and similarly throughout.

© A. Authors.

In tandem, Assumptions 1 and 2 guarantee that $H_\phi \neq G$, and in particular that $\texttt{supp}(H_\phi) \supsetneq \texttt{supp}(G)$, for each ϕ. Hence, by using H to model the random variable Y, we are introducing *intentional model misspecification*; we intentionally do inference with a model that has inflated support. In order construct a valid approximate inference procedure, we require the pair (F_θ, H_ϕ) to satisfy the *dominated likelihood* assumption

$$h_\phi(y) \leq f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi. \qquad \textit{(Assumption 3)} \tag{1.0.3}$$

In particular, $\Phi \subseteq \Theta$, so we may substitute parameters from Φ into Θ. Since we do inference with the dominated approximate likelihood H to get approximate inference for the target likelihood F, we refer to our strategy as *dominated likelihood approximation* (DLA).

2. Methods

Here we detail and justify doing approximate inference on a target observation model F via inference on a dominated observation model H with broader support. In the frequentist paradigm (Sec. 2.1), we can substitute a parameter learned with H by maximum likelihood into F. In the Bayesian paradigm (Sec. 2.2), we can treat the posterior using H as an approximate posterior using F. We can also treat a variational approximate posterior using H as a doubly approximate posterior using F.

2.1. Maximum Likelihood

Although the true distribution governing Y is G, we can define the **quasi-maximum likelihood estimator** (QMLE) (White, 1982) for both our target observation model F and our support-broadened observation model H via:

$$\widehat{\theta}_n \triangleq \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{\sum_{i=1}^n \log f_\theta(y_i)}{n}, \qquad \widehat{\phi}_n \triangleq \underset{\phi \in \Phi}{\operatorname{argmax}} \frac{\sum_{i=1}^n \log h_\phi(y_i)}{n}$$

In the following, we justify the utility of both (1) $H_{\widehat{\phi}_n}$ and (2) $F_{\widehat{\phi}_n}$ as reasonable models for the random variable Y, even though ϕ indexes a distribution H that has artificially and incorrectly broadened support, i.e. $\texttt{supp}(H) \supsetneq \texttt{supp}(G)$. Model (2) is especially useful; we do inference (QMLE) with a model H that has artificially inflated support, and then substitute the learned parameter into the model F which generates observations with the correct support.

1. Justification for $H_{\widehat{\phi}_n}$. Here we sketch the justification; see Sec. B for details. The justification follows from known properties of the QMLE (White, 1982). By definition, we have that the quasi-MLE $\widehat{\phi}_n$ maximizes $L_n(h) \triangleq \frac{1}{n}\sum_{i=1}^n \log h_\phi(y_i)$. By the Strong Law of Large Numbers, $\lim_{n\to\infty} L_n(h) = \mathbb{E}_G[\log h_\phi(Y)]$ $[G]$, where the symbol $[G]$ means G almost surely. Hence, with probability 1, $\widehat{\phi}_n$ asymptotically maximizes $\mathbb{E}_G[\log h_\phi(Y)]$, and so asymptotically minimizes $\texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$. In our case, $\texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$ is well-defined even though $\texttt{supp}(H) \supsetneq \texttt{supp}(G)$, because the KL-divergence assigns zero probability mass to any points in $\texttt{supp}(H) \setminus \texttt{supp}(G)$. Thus, performing maximum likelihood with the support-broadened model H is reasonable because the procedure asymptotically minimizes $\texttt{KL}\big[G \;\big|\big|\; H_\phi\big]$.

2. Justification for $F_{\widehat{\phi}_n}$. We have

$$\begin{aligned} h_\phi(y) &\leq f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi && \text{Assumption 3} \\ \implies \mathbb{E}_G[\log h_\phi(Y)] &< \mathbb{E}_G[\log f_\phi(Y)] && \text{Monotonicity, Assumption 1} \\ \iff \mathbb{E}_G[\log g(Y)] - \mathbb{E}_G[\log f(Y)] &< \mathbb{E}_G[\log g(Y)] - \mathbb{E}[\log h_\phi(Y)] && \text{algebra} \\ \iff \mathtt{KL}\big[G \,\big|\big|\, F_\phi\big] &< \mathtt{KL}\big[G \,\big|\big|\, H_\phi\big] && \text{def. KL} \end{aligned} \tag{2.1.1}$$

where for simplicity we have assumed that G has density g. Now by item 1, we have that $\widehat{\phi}_n$ asymptotically minimizes $\mathtt{KL}\big[G \,\big|\big|\, H_\phi\big]$, which by Eq. (2.1.1) is an upper bound on $\mathtt{KL}\big[G \,\big|\big|\, F_\phi\big]$. Hence, substituting the quasi-MLE $\widehat{\phi}_n$ from family H into family F to obtain model $F_{\widehat{\phi}_n}$ can be justified since $\widehat{\phi}_n$ is the parameter in Φ which (asymptotically) minimizes an upper bound on $\mathtt{KL}\big[G \,\big|\big|\, F\big]$.

Discrepancy induced by the dominated likelihood approximation. From Eq. (2.1.1), we see that the discrepancy induced by support-broadening – more specifically, the asymptotic discrepancy between the objective functions whose optimizations produce the approximate (DLA) model $F_{\widehat{\phi}_n}$ rather than the target (quasi) MLE model $F_{\widehat{\theta}_n}$ – is given by

$$\mathcal{D}(F_\phi, H_\phi) \triangleq \mathtt{KL}\big[G \,\big|\big|\, H_\phi\big] - \mathtt{KL}\big[G \,\big|\big|\, F_\phi\big] = \mathbb{E}_G[\log f_\phi(Y)] - \mathbb{E}_G[\log h_\phi(Y)] \geq 0 \tag{2.1.2}$$

If the target model is well-specified (i.e. $G = F$), then $\mathcal{D}(F_\phi, H_\phi) = \mathtt{KL}\big[F_\phi \,\big|\big|\, H_\phi\big]$. If F_ϕ is a truncation of H_ϕ (i.e. $f_\phi = h_\phi / Z_\phi$, where the normalizing constant Z_ϕ gives the probability mass that H_ϕ assigns to some truncation region), then

$$\mathcal{D}(F_\phi, H_\phi) = \sum_{i=1}^{n} \mathcal{D}_i(F_\phi, H_\phi) = \sum_{i=1}^{n} (-\log Z_{i,\phi}) \tag{2.1.3}$$

where $\mathcal{D}_i(F_\phi, H_\phi) \triangleq -\log Z_{i,\phi}$ and $Z_{i,\phi}$ is the normalizing constant for the ith observation. That is, $Z_{i,\phi}$ is the probability mass assigned by the approximating model H_ϕ to the event that the i-th observation lies within the truncation region. Interestingly, Eq. (2.1.3) reveals a *win-win* scenario: the greater the modeling benefit of truncation, the lower the cost of DLA.

2.2. Bayesian inference

Justification for approximate Bayesian inference via dominated likelihoods. Given a prior distribution π on Φ, we obtain the following marginal density relationship from Assumption 3:

$$p_F(y) \triangleq \int_\Phi f_\phi(y)\, \pi(d\phi) \geq \int_\Phi h_\phi(y)\, \pi(d\phi) \triangleq p_H(y)$$

where we have used the same prior π on both Φ and Θ, using the implication from Assumption 3 that $\Phi \subseteq \Theta$. Hence, for any probability distribution Q on Φ within some chosen family $\mathcal{Q}$, we have

$$\log p_F(y) \geq \log p_H(y) \geq \mathtt{ELBO}_H(Q) \tag{2.2.1}$$

where

$$\mathtt{ELBO}_H(Q) \triangleq \mathbb{E}_Q[\log h(y \mid \phi)] - \mathtt{KL}\big[Q \,\big|\big|\, \pi\big]$$

is the *evidence lower bound*, a traditional lower bound on the log marginal likelihood of a Bayesian model (Blei et al., 2017). In other words, variational inference using the support-broadened likelihood H, which finds $Q \in \mathcal{Q}$ to maximize $\texttt{ELBO}_H(Q)$, not only maximizes an upper bound on (the logarithm of) the marginal density when using the *support-broadened* observation model p_H, but also when using the *target* observation model p_F. This justifies DLA in the context of variational Bayesian inference. Moreover, as is well-known, when Φ is unconstrained, $\texttt{ELBO}_H(Q)$ is optimized by the true posterior π_H^n, defined as the posterior distribution on Φ after observing $Y_1, \ldots, Y_n$ when assuming the observation model H. Thus, Eq. (2.2.1) also says that ordinary Bayesian inference (such as MCMC) for computing the posterior on Φ using H can be seen as producing the probability distribution on Φ which maximizes a lower bound on p_F.

Perturbation to posterior concentration. Define π_D^n as the posterior distribution on Φ after observing $Y_1, \ldots, Y_n$ when assuming a generic observation model D with density d. Then $\Phi_0 \subset \Phi$ is called an *asymptotic carrier* for Φ if for any open set U containing Φ_0, $\lim_{n\to\infty} \pi_D^n(U) = 1\,[G]$. Berk (1966) shows that under mild conditions, an *asymptotic carrier* for Φ is given by

$$\Phi_0 = \{\phi^* \in \Phi : \phi^* = \operatorname*{argmax}_{\phi\in\Phi} \mathbb{E}_G[\log d(Y \mid \phi)]\} \tag{2.2.2}$$

Since $\mathbb{E}_G[\log d(Y|\phi)]$ is maximized when $D = G$, the set Φ_0 is precisely those $\phi \in \Phi$ that minimize $\texttt{KL}\big[G \,\big|\big|\, D\big]$. Therefore, we can also write

$$\Phi_0 = \{\phi \in \Phi : \phi = \text{argmin}_{\phi\in\Phi} \texttt{KL}\big[G \,\big|\big|\, D\big]\}. \tag{2.2.3}$$

So as n increases, the Bayesian posterior concentrates its support on regions of parameter space given by the limiting values of the QMLE. Thus, when choosing to use a dominated observation model $D = H_\phi$, we can apply the argument of Eq. (2.1.1) to find that the posterior π_H^n concentrates on regions of parameter space Φ that minimize an upper bound on $\texttt{KL}\big[G \,\big|\big|\, F_\phi\big]$. This observation justifies the continued relevance in the Bayesian setting of the discrepancy $\mathcal{D}(F, H)$ in Eq. (2.1.2) to quantify the error induced by DLA; $\mathcal{D}(F, H)$ gives the amount of perturbation that DLA imposes upon the objective function which is maximized by the asymptotic carrier.

3. Applications

3.1. Truncating a mixture of Gaussian distributions

Geolocation data often have constrained support due to geographic constraints, political boundaries, etc. While using truncated distributions is possible, fitting these in practice may not be easy. Consider a bivariate Normal distribution with truncated support to one rectangular region $\mathcal{Y}$ with bounds a, b such that $y_1 \in [a_1, b_1], y_2 \in [a_2, b_2]$. Given a fixed dataset, the ML estimate of Gaussian parameters μ, Σ requires iterative optimization. In contrast, untruncated Gaussian parameters can be estimated in closed-form.

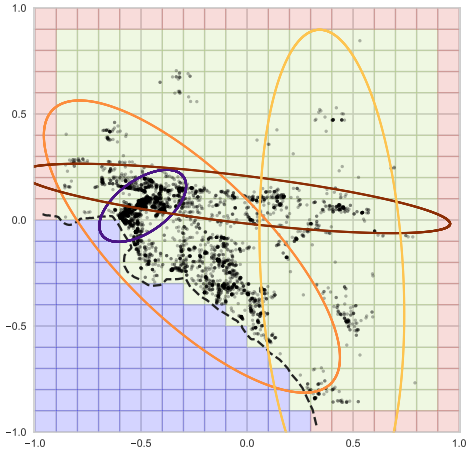
We can apply our DLA framework to pursue density modeling of human geolocation data. We use Cho et al. (2011)'s open-access dataset of user "check-ins" to the Gowalla website, focusing on Southern California as in Lichman and Smyth (2014). Our training set is illustrated in Fig. 1. Each observation y_i represents a vector of latitude and longitude coordinates. Naturally, checkins happen on land (green) but not water (blue). We show a coarse 20x20 grid of rectangular regions. We can mark each region as either in-bounds (green) or out-of-bounds (blue or red).

We pursue mixtures of union-of-rectangle truncated Gaussians f, via tractable untruncated h:

$$f(y) = \sum_{k=1}^{K} \pi_k \text{TruncNormPDF}_{a,b}(y|\mu_k, \Sigma_k)$$
$$h(y) = \sum_{k=1}^{K} \pi_k \text{NormPDF}(y|\mu_k, \Sigma_k)$$

We know $f(y|\phi) > h(y|\phi)$ for all parameters ϕ (all valid frequencies π, means μ, and covariances Σ) as well as all geolocations y in our truncated region $\mathcal{Y}$, because by construction $f(y|\phi) = \frac{1}{Z_{a,b}(\phi)} h(y|\phi)[y \in \mathcal{Y}]$, where $Z_{a,b}(\phi) < 1$ is the probability mass that h allocates to the union-of-rectangular regions defined by bounds a, b (e.g. green rectangles in Fig. 1).

After fitting parameters ϕ via VB coordinate ascent for unconstrained GMMs, we assess the likelihood of heldout user checkins using unconstrained h as well as truncated f via parameter substitution. Fig. 1 demonstrates that our approach – fit with easy h, then plug-in to f – yields higher test likelihoods, because at test-time we better account for the support of the data.



Training	Inference	Test log like
K=4 GMM-VB	h : Normal	0.3096
as above	f : TruncNormal	0.3506

Figure 1: **Application: GMMs for geolocations in southern CA**. Ideal model f truncates to the union of green rectangles (land area). Tractable model h allocates mass to water (blue) or out-of-bounds (red). Ellipses show the 99% high-density-areas of 4 Gaussian clusters fit to data.

3.2. Categorical Generalized Linear Models

Bayesian inference with categorical generalized linear models (GLMs) is surprisingly difficult to scale to large datasets (Wojnowicz et al., 2022). For instance, coordinate ascent variational inference (CAVI) (Blei et al., 2017) with multi-logit (a.k.a. softmax) regression faces an expected log-sum-exp term, a notorious blocker to closed-form CAVI (Braun and McAuliffe, 2010; Wang and Blei, 2013). The problematic term is distinct to the multi-class case; it does not appear for binary regression models (e.g. logistic regression). This motivates the question: is it possible to use binary regression models to approximate the posterior from a categorical regression model?

Motivated by this question, Wojnowicz et al. (2022) define *categorical-from-binary* (CB) models, a new class of categorical GLMs that are constructed from independent binary (IB) models (products of binary-outcome regression models, such as logistic regressions). Interpreting categorical variables as one-hot encoded vectors, IB models have broadened support (Assumption 1) relative to CB models:

$$G = \text{true distn}, \qquad F_\theta = \text{CB model} \qquad \text{one-hot space} \subset \text{K-bit space}$$
$$H_\phi = \text{IB model} \qquad \text{K-bit space}$$

The authors construct two variants of CB models, CBC and CBM models (see Sec. D), and both variants dominate IB models (in the strict sense of Assumption 3'; see Eq. (A.0.1)). Hence, the DLA framework applies.

Here we focus on CBC models (Eq. (D.0.2)), which are truncated IB models. When we approximate a CBC likelihood with an IB likelihood (representing categories as one-hot vectors), Eq. (2.1.3) reveals that the asymptotic discrepancy in the objective function for maximum likelihood is given by

$$\mathcal{D}(F_{\theta=\phi}, H_\phi) = \sum_{i=1}^{n}(-\log Z_i), \quad \text{where} \quad Z_i \triangleq \sum_{\ell=1}^{K} C(\eta_{i\ell}) \prod_{j\neq\ell}(1 - C(\eta_{ij})) \tag{3.2.1}$$

Here, Z_i is the probability mass assigned by the IB model to the event that the i-th observation is one-hot encoded.[3] Fig. 2 reveals that, as predicted by our framework, the quality of the IB approximation improves as the IB model assigns higher probability to one-hot space. In this way, we can assess the quality of the approximation to the CBC model without ever fitting it.

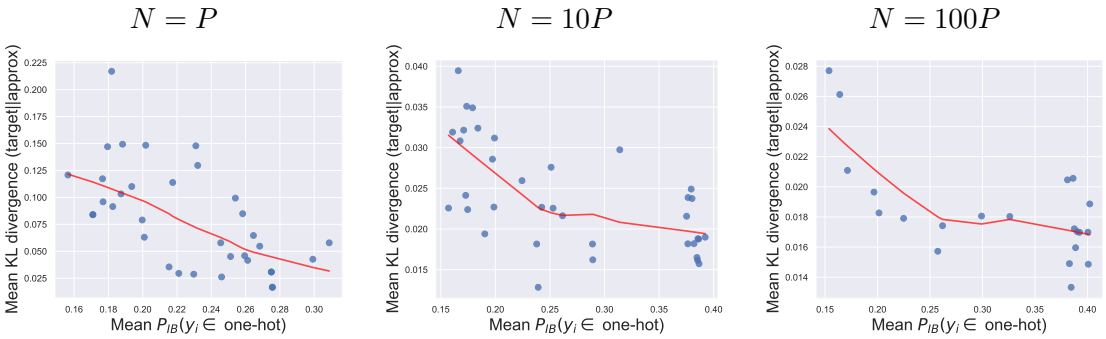


Figure 2: **Application: When do independent binary models give good approximations to categorical GLMs?** Each panel summarizes results on multiple simulated categorical regression datasets where the number of observations, N, is expressed as a multiple of the number of parameters P. The y-axis shows the empirical mean KL divergence from an independent binary (IB) approximation to a categorical-from-binary (CB) target model. The x-axis gives the empirical mean probability that the IB model assigns to one-hot encoded space.

4. Discussion

Although intractable normalizing constants are a fundamental barrier to statistical inference, most approximation procedures provide solutions for particular models and/or particular inference approaches (e.g. Depraetere and Vandebroek (2017); Atchadé et al. (2013)). Our support-broadening procedure applies in principle to any likelihood with an intractable normalizing constant and to most common inference procedures, so long as a dominated likelihood can be found. Although specialized approaches can likely provide better approximations in specific settings, our procedure yields a simple tool that can be used for quickly and easily obtaining approximate inference in a wide range of settings.

[3] When the IB model assigns all its probability mass to the event that the i-th observation is one-hot encoded, i.e. $Z_i = 1$, we have $-\log Z_i = 0$, and so the discrepancy for the i-th observation attains the minimum value of $\mathcal{D}_i = 0$. When the IB model assigns none of its probability mass to the event that the i-th observation is one-hot encoded, i.e. $Z_i = 0$, we have $-\log Z_i = \infty$, and so the discrepancy for the i-th observation attains attains the maximum value of $\mathcal{D}_i = \infty$.

References

Yves F Atchadé, Nicolas Lartillot, and Christian Robert. Bayesian computation for statistical models with intractable normalizing constants. 2013.

R.R. Bahadur. Some limit theorems in statistics. *SIAM*, 1971.

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

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.

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.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.

Nicolas Depraetere and Martina Vandebroek. A comparison of variational approximations for fast inference in mixed logit models. *Computational Statistics*, 32:93–125, 2017.

Robert Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, pages 633–643, 1969.

J Kiefer and J Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, pages 887–906, 1956.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 2017.

Moshe Lichman and Padhraic Smyth. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

M.R. Mickey. Test criteria for pearson type iii distributions. *Aerospace Research Laboratories*, pages 40–41, 1963.

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

Halbert White. Nonlinear regression on cross-section data. *Econometrica: Journal of the econometric society*, pages 721–746, 1980.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

Michael T Wojnowicz, Shuchin Aeron, Eric L Miller, and Michael Hughes. Easy variational inference for categorical models via an independent binary approximation. In *International Conference on Machine Learning*, pages 23857–23896. PMLR, 2022.

Appendix A. Strict Dominated Likelihood Assumption

Note that equality cannot be attained in Eq. (1.0.3) for all $y \in \mathcal{Y}$; by Assumption 1, for any parameters $(\phi, \theta) \in \Phi \times \Theta$, there must be some $A \in \mathcal{F}$ such that $0 < H_\phi(A) < F_\theta(A)$. Hence, in the most straightforward application of the framework, one imposes a *strict dominated likelihood* assumption

$$h_\phi(y) < f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi. \qquad \textit{(Assumption 3')} \tag{A.0.1}$$

Indeed, since probability densities must satisfy $\int_{\mathcal{Y}} f_\theta(y)\, \mu(dy) = \int_{\mathcal{Y}_*} h_\phi(y)\, \nu(dy) = 1$, Assumption 3' implies Assumption 1.

Appendix B. Quasi Maximum Likelihood Estimation

B.1. Existence of the quasi maximum likelihood estimator (QMLE)

Here we state and verify the assumptions given by (White, 1982) pertaining to the existence of the QMLE $\hat{\phi}_n$. These assumptions are standard in statistical theory when establishing properties such as consistency of estimators.

Assumption A1: The independent random variables Y_i for $i \in \{1, \ldots, N\}$ have a common joint distribution function G on Ω, a measurable Euclidean space, with Radon-Nikodym derivative $g = \frac{dG}{d\nu}$

Verification of A1: The first assumption simply defines the space Ω and $g = \frac{dG}{d\nu}$ as the Radon-Nikodym derivative of a distribution function G on Ω. In most scenarios, we do not know the underlying true distribution G. In Section 1, the setting is a measurable space $(\mathcal{Y}, \mathcal{H})$ where G is a probability distribution on $\mathcal{H}$ with Radon-Nikodym derivative $g = \frac{dG}{d\nu}$. Thus A1 is met. When we have access to the underlying true distribution such as when generating data for the categorical GLM experiments in Section 3, this assumption is verified as any probability mass function gives a Radon-Nikodym derivative with respect to the counting measure.

Assumption A2: The family of distribution functions $\{F_\theta : \theta \in \Theta\}$ and $\{H_\phi : \phi \in \Phi\}$ has Radon-Nikodym derivatives $f_\theta = \frac{dF_\theta}{d\nu}$ and $g_\phi = \frac{dH_\phi}{d\nu}$ which are measurable in x for every $\theta \in \Theta$ and $\phi \in \Phi$ and continuous in θ and ϕ for every $x \in \Omega$. The parameter spaces Θ and Φ are compact subsets of Euclidean space.

Verification of A2: The second assumption requires the existence of Radon-Nikodym derivatives for the distributions F_θ and H_ψ and that the Radon-Nikodym derivatives were continuous with respect their parameters for all $y \in \mathcal{Y}$. We assumed this was the case in Section 1. This is met in practice and in our experiments: any probability measure that is absolutely continuous (has a probability density function) or discrete (has a probability mass function) automatically has a Radon-Nikodym derivative , with respect to Lebesgue and counting measure, respectively. This Radon-Nikodym derivative is automatically measurable.

This assumption also requires that the parameter space Θ is compact, which on its face does not hold in practice. For example, we generally do not restrict the space of the mean parameter $\mu \in \mathbb{R}$ of a normal distribution $N(\mu, \sigma^2)$ when estimating it using maximum likelihood; yet $\mathbb{R}$ is not compact. To address this technical issue, it is standard practice to *compactify* the parameter spaces Θ and Ψ

if necessary, wherein points are added to the spaces so as to make them compact. See (Bahadur, 1971) and (Kiefer and Wolfowitz, 1956) for a discussion.

Theorem (White, 1982) - Existence of a QMLE: Given assumptions A1 and A2, there exists a measurable QMLE.

Verification of Theorem 2.1: Since assumptions A1 and A2 are met, we conclude that the QMLEs $\hat{\theta}_n$ and $\hat{\phi}_n$ for both our observation model F and our support-broadened observation model H from Sec. 2.1 exist.

B.2. Convergence of QMLE and Justification for $H_{\hat{\phi}_n}$

Here we detail the argument laid out in Section 2.1 to justify $H_{\hat{\phi}_n}$ as a reasonable model despite the support-inflation. The argument in three steps is as follows:

1. $\hat{\phi}_n$ maximizes $L_n(h) \triangleq \frac{1}{n}\sum_{i=1}^{n} \log h_\phi(y_i)$ is the definition of the QMLE
2. $\lim_{n\to\infty} L_n(h) = E_G[\log h_\phi(Y)]$ is known as *Mickey's Theorem* (Theorem 2 in (Jennrich, 1969) and p. 40 in (Mickey, 1963) contain two proofs) and is a Law of Large Numbers-type result

   Theorem (Jennrich, 1969) - Convergence of Quasi-Likelihoods: Let q be a function on $\Omega \times \Theta$ such that $q(z,\theta)$ is a continuous function of θ for each z and a measurable function of z for each θ. Suppose also that $|q(z,\theta)| \leq m(z)$ for all z and θ, where m is integrable with respect to distribution G on Ω. If $Z_1, Z_2, \ldots$ is a random sample from G, then $Q_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} q(z_i,\theta)$ converges to $Q(\theta) \triangleq \int q(z,\theta)dG(z)$ uniformly for all $\theta \in \Theta$ and almost every sequence $\{Z_i\}$.

   Verification of Theorem (Jennrich, 1969) - Convergence of Quasi-Likelihoods:

   The only item to verify here is the bounded condition by a function that is integrable with respect to the distribution G, which we often do not know. In our application to categorical GLMs via an independent binary model, we show that the summands in the quasi log-likelihood $|\log(p_{IB(y,B)})|$ are integrable with respect to Lebesgue:

$$
\begin{aligned}
|\log(p_{IB}(y,B))| &= |\sum_{k=1}^{K} \log[H(x^T\beta_k)(1-H(x^T\beta_k))]| \\
&\leq \sum_{k=1}^{K} |\log[H(x^T\beta_k)(1-H(x^T\beta_k))]| \\
&\leq \sum_{k=1}^{K} |\log[H(x^T\beta_k)]|
\end{aligned}
$$

   Now, the function $H(\cdot) \in (0,1)$ so that $|\log[H(\cdot)]| \in (0,\infty)$. This function is bounded above by 1 on the interval $(e^{-1},1)$ so we are only concerned with bounding it by an integrable function on the interval $(0,e^{-1})$. This is accomplished with the function $|\log[H^2(\cdot)]|$, which bounds $|\log[H(\cdot)]|$ from above and is integrable with respect to Lebesgue despite growing rapidly close to zero. Thus each term in the above summand is integrable and therefore $|\log(p_{IB}(y,B))|$ is integrable with respect to Lebesgue

3. With probability 1, $\hat{\phi}_n$ asymptotically maximizes $E_G[\log h_\phi(Y)]$ and so asymptotically minimizes $\texttt{KL}\big[G \,\big|\big|\, H_\phi\big]$ is *Lemma 2.2* from (White, 1980)

Theorem (White, 1980) - Convergence of QMLE to a Minimizer of Cross-Entropy: Let $Q_n(\omega, \theta)$ be a measurable function on a measurable space Ω and for each $\omega \in \Omega$ a continuous function on a compact set Θ. Furthermore, let $\bar{Q}_n(\theta)$ be a continuous function on Θ such that $\bar{Q}_n(\theta)$ has a unique minimum $\theta_0 \in \Theta$ for all n. Then, there exists a measurable function $\hat{\theta}_n(\omega)$ such that $Q_n(\omega, \hat{\theta}_n(\omega)) = \inf_{\theta \in \Theta} Q_n(\omega, \theta)$ for all $\omega \in \Omega$. Furthermore, if $|Q_n(\omega, \theta) - \bar{Q}_n(\theta)| \to 0$ a.e. uniformly for all $\theta \in \Theta$ and if for all n we have $\bar{Q}_n(\theta)$ has a minimum at θ_0 which is identifiably unique, then $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$

Verification of (White, 1980) - Convergence of QMLE to a Minimizer of Cross-Entropy:

The space defined in Section 2.1 $(\mathcal{Y}, \mathcal{F})$ is a measurable space and the distributions F_θ, H_ϕ, and $\mathcal{G}$ have Radon-Nikodym derivatives that are automatically measurable. The function $Q_n(\omega, \theta)$ in our setting is a negative (quasi) log-likelihood of the misspecified and support-broadened model defined in Section 2.1: $Q_n(\omega, \theta) = -L_n(h) = -\frac{1}{n}\sum_{i=1}^n \log h_\phi(y_i)$. This function is a sum of compositions of measurable functions, hence measurable. Hence, we conclude that there exists an measurable function $\hat{\phi}_n(y)$, that achieves the infimum lower bound and minimizes $-L_n(h)$ i.e. maximizes $L_n(h)$. This is a restatement of the existence of the QMLE laid out in the previous section verify the assumptions A1 and A2 in (White, 1980). Now, by the above theorem in (Jennrich, 1969):

$$\lim_{n\to\infty} L_n(h) = E_G[\log h_\phi(Y)]$$

Furthermore, the convergence is guaranteed to be uniform $\forall \phi \in \Phi$. Note, that this limit is the negative cross-entropy of H_ϕ relative to G: $H(G, H_\phi) = -E_G[\log h_\phi(Y)]$. Clearly, maximizing the negative cross-entropy minimizes the cross-entropy, which in turn minimizes the $\texttt{KL}\big[G \,\big|\big|\, H_\phi\big]$ via the relation: $H(G, H_\phi) = H(G) + \texttt{KL}\big[G \,\big|\big|\, H_\phi\big]$. Therefore, if the cross-entropy has an identifiably unique minimum $\phi^* \triangleq \operatorname{argmin}_{\phi\in\Phi} E_G[\log h_\phi(Y)]$, we have convergence of the QMLE to a minimizer:

$$\hat{\phi}_n \xrightarrow{a.s.} \phi^*$$

It is shown in (Wojnowicz et al., 2022) that the CB models in are non-identifiable at least in the intercepts-only setting. One route to rectify the issue of non-identifiability is a technical topological argument that involves passing to the quotient topology of the parameter wherein the set of identifiable parameters are gathered into an equivalence class, the maximum likelihood estimator is indeed consistent estimator for the true parameter of interest. However, (Wojnowicz et al., 2022) also showed that the IB models are *globally identifiable*, at least in the intercepts-only setting. Thus, since the IB model is the support-inflated model on which we do QMLE, the identifiability condition on the minimizer ϕ^* is satisfied.

B.3. Bayesian Inference and Perturbation to Posterior Concentration

The justification for $F_{\hat{\phi}_n}$ laid out in the second half of Section 2.2 is based on a classic Bayesian consistency result laid out in (Berk, 1966). This result guarantees a Bayesian formulation of the

$\texttt{KL}\big[G \;||\; F_\phi\big]$ upper-bound minimization and asymptotic discrepancy results in 2.1: when using a dominated observation model the posterior π_H^n concentrates on regions of parameter space that minimize $\texttt{KL}\big[G \;||\; H_\phi\big]$. Recall the definition of the asymptotic carrier: $\Phi_0 \subset \Phi$ is called an *asymptotic carrier* for Φ if for any open set U containing Φ_0, $\lim_{n\to\infty} \pi_D^n(U) = 1\,[G]$

Assumption B1: $f(x|\theta)$ is measurable jointly in x and θ; for almost every x, $f(x|\cdot)$ is continous in θ, at all $\theta \in \Theta$

Interpretation and Verification of B1: This is a standard measurability and continuity assumption that we have assumed from the outset in Section 1

Assumption B2: For all $\theta \in \Theta$, $F\{x : f(x|\theta) > 0\} = 1$

Interpretation and Verification of B2: (Berk, 1966) states that this assumption avoids situations where we obtain realizations of the random variables $\{X_i\}$ for which the posterior may be undefined on certain subsets $A \subset \Theta$ and that this is not a restrictive assumption. In the case where it does not hold, it is possible to rectify this by constructing a sequence of sets A_i such that $\bigcup_{i\in\mathcal{I}} A_i = A$ where $F[\lim_{k\to\infty} P_k A_i = 0] = 1$

Assumption B3: For every $\theta \in \Theta$, there is an open neighborhood U of θ such that the expected value over U of $||H(X|\cdot)||_\infty$ is finite, where $H(X|\cdot)$ is the logarithm of a function of the density $f(x|\theta)$

Interpretation and Verification of B3: This together with B4 are boundedness conditions so that the dominated convergence theorem can be applied in the proof of the theorem below

Assumption B4: There is a positive integer p such that for every real number r there is a compact subset D of Θ such that the expected value over D of $\sup \bar{H}_p$ is less than or equal to r, where $\bar{H}_p = \frac{1}{p}\sum_{i=1}^{p} H(X_i|\theta)$

Interpretation and Verification of B4: Again, this together with B3 are boundendess conditions so that the dominated convergence theorem can be applied. In the case where $H(X|\cdot) = \log f(x|\theta)$, this amounts to a finite second moment condition $E[X^2] < \infty$

Theorem (Berk, 1966) - Concentration of Posterior under Misspecification: Suppose a model for the random variables $\{X_i\}$ specifies they are i.i.d. with one of densities $f(\cdot|\theta)$, where the range, Θ, is a Borel subset of a complete separable metric space and $f(\cdot|\theta)$ are densities with respect to a fixed σ-finite measure on range X. Let P be a prior distribution on the Borel subsets of Θ and let P_k be the posterior distribution of θ given $X_1, \ldots, X_k$. If the $\{X_i\}$ are in fact distributed according to a distribution F, and assumptions (B1)-(B4) hold, then P_k is almost surely $[F]$ asymptotically carried on the asymptotic carrier.

Appendix C. Supplementary material on Union-of-Rectangle Truncated GMMs

We define a Gaussian mixture model truncated to a union of rectangular regions. Assume the entire space $\mathbb{R}^2$ is divided into an infinite grid of non-overlapping rectangles. A finite number of rectangles R is selected as the truncation region. Each one is indexed by integer r, and has lower bounds $a_r = [a_{1r}, a_{2r}]$ and upper bound $b_r = [b_{1r}, b_{2r}]$ such that an observed 2-dimensional vector y is in the rectangle if $y_1 \in [a_{1r}, b_{1r}]$ and $y_2 \in [a_{2r}, b_{2r}]$. We represent the entire truncation region as the union of the selected rectangles, with bounds $a = \{a_1, \ldots a_R\}$ and $b = \{b_1, b_2, \ldots b_R\}$.

The probability mass Z_r that a GMM allocates to one rectangle r can be defined as

$$Z_r = \int_{y \in [a_r, b_r]} \sum_{k=1}^{K} \pi_k \text{NormPDF}(y|\mu_k, \Sigma_k) dy \tag{C.0.1}$$

$$= \sum_{k=1}^{K} \pi_k \int_{y \in [a_r, b_r]} \text{NormPDF}(y|\mu_k, \Sigma_k) dy \tag{C.0.2}$$

$$= \sum_{k=1}^{K} \pi_k (F_k(b_{1r}, b_{2r}) - F_k(b_{1r}, a_{2r}) - F_k(a_{1r}, b_{2r}) + F_k(a_{1r}, a_{2r})) \tag{C.0.3}$$

where F_k is the multivariate normal CDF under mean μ_k and covariance Σ_k. Each CDF evaluation computes the mass assigned to the rectangle whose lower corner is $(-\infty, -\infty)$ and whose upper right corner has the provided coordinates. The addition and subtraction handles taking the differences of such larger rectangles to compute the area of the desired rectangle. CDF evaluations for fixed parameters are rapid, and can be cached if needed for multiple computations (multiple rectangles that share some boundary corners).

The overall probability mass assigned to the union of R rectangles is simply

$$Z = \sum_{r=1}^{R} Z_r \tag{C.0.4}$$

As long as the rectangles do not cover all of $\mathbb{R}^2$, we know that $Z < 1$.

C.1. Gowalla data preprocessing

We extract geolocation records from the Gowalla dataset that fall within a rectangular area of southern California, with longitude in (-118.825, -116.675) and latitude in (33.14, 34.86). We then keep only records for individual users that are spaced at least 6 hours apart (to avoid too much time dependence), then only keep users with at least 20 observations. We divide data by user so that each user's records belong to one of train/valid/test. 80% of users are allocated to training (at random), the rest split equally between validation and test. We have 67925 records in train, 8375 records in the test set. Validation set was not used here. Very few records had locations over water, likely due to boating, geolocation errors, or people visiting some of the islands off the coast. We removed these for simplicity, but note that our union of rectangle approach could easily handle islands.

Appendix D. Supplementary material on Categorical-from-Binary (CB) models

Categorical-from-binary (CB) models (Wojnowicz et al., 2022) are categorical GLMs for which independent binary (IB) models (products of binary-outcome regression models) naturally provide a dominated likelihood (in the strict sense of Assumption 3').

The the likelihood for K-bit observation $\widehat{\boldsymbol{y}}_i = (\widehat{y}_{i1}, \ldots \widehat{y}_{iK}) \in \{0,1\}^K$ under an *independent binary* (IB) model is

$$h_{\text{IB}}(\widehat{\boldsymbol{y}}_i \mid \widehat{\boldsymbol{B}}) = \prod_{k=1}^{K} C(\widehat{\eta}_{ik})^{\widehat{y}_{ik}} \left(1 - C(\widehat{\eta}_{ik})\right)^{1-\widehat{y}_{ik}}, \tag{D.0.1}$$

where C is an arbitrary cumulative distribution function, where each linear predictor $\widehat{\eta}_{ik} = \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}_k$ is formed from known covariates $\boldsymbol{x}_i \in \mathbb{R}^M$ and unknown parameters $\widehat{\boldsymbol{\beta}}_k \in \mathbb{R}^M$, and 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.

A *categorical-from-binary-via-conditioning* (CBC) model assigns probabilities to categorical observations $y_i \in \{1, \ldots, K\}$ by conditioning the IB model on the event that the vector has exactly one positive entry:

$$f_{\text{CBC}}(y_i{=}k \mid \boldsymbol{B}) = \frac{C(\eta_{ik}) \prod_{j \neq k} (1 - C(\eta_{ij}))}{\sum_{\ell=1}^{K} C(\eta_{i\ell}) \prod_{j \neq \ell} (1 - C(\eta_{ij}))} \tag{D.0.2}$$

for categories $k = 1, ..., K$. Here, we distinguish parameters $\boldsymbol{B}$ and observations y_i from the IB case by dropping the hat accent.

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

$$f_{\text{CBM}}(y_i = k \mid \boldsymbol{B}) = \frac{C(\eta_{ik})}{\sum_{\ell=1}^{K} C(\eta_{i\ell})}. \tag{D.0.3}$$

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

D.1. Experiment details

We simulated categorical regression datasets following the procedure of Wojnowicz et al. (2022, Sec G.1). In particular, we formed three collections of 36 simulated datasets generated from a softmax categorical GLM with $K = 5$ categorical outcomes and $M = 5$ covariates (which along with the intercept gives $P = K(M + 1) = 30$ parameters). Each collection of datasets was defined by the number of observations (or examples), N, given as a multiple of the number of predictors $N \in \{1P, 10P, 100P\}$. For each collection, datasets were generated with different levels of category predictability (Wojnowicz et al., 2022, Sec. G.2).

The target model for inference F was a CBC-Probit likelihood with an isotropic Gaussian prior. The dominated approximation H was given by an IB-Probit model (the product of binary probit regressions). Inference for H was performed by a lightweight coordinate ascent variational inference (CAVI) algorithm given in Wojnowicz et al. (2022, Sec. D). Inference for F was performed by automatic differentiation variational inference (ADVI) (Kucukelbir et al., 2017). In both cases, the true parameter was approximated by the variational posterior expectation.

.