Approximate inference by broadening the support of the likelihood

Michael T. Wojnowicz [1 3], Martin Buck [1 2], Michael C. Hughes [1 3]

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

Overview

We present a framework for approximate statistical inference on a target observation model F via inference on an observation model H with broader support which gives relatively easy and efficient inference.

Setup

Suppose we observe the random variables $Y_i \overset{\text{i.i.d.}}{\sim} G$ for $i = 1, \dots, n$, where G is an unknown probability distribution on $\mathcal{Y}$.

We have

$Y_i \overset{\text{i.i.d.}}{\sim} G$	True Data Generating Process
$Y_i \overset{\text{i.i.d.}}{\sim} F_\theta, \quad \theta \in \Theta$	Target Model
$Y_i \overset{\text{i.i.d.}}{\sim} H_\phi, \quad \phi \in \Phi$	Approximation Model

We assume each F_θ has density f_θ w.r.t. some σ-finite measure μ on the measurable space $(\mathcal{Y}, \mathcal{F})$. We also assume each H_ϕ has density h_ϕ w.r.t. some σ-finite measure ν on a space $(\mathcal{Y}*, \mathcal{F}*)$ where $\mathcal{Y}* \supsetneq \mathcal{Y}, \mathcal{F}* \supsetneq \mathcal{F}$. Finally, we assume these densities are continuous w.r.t θ and ϕ, respectively, for each $y \in \mathcal{Y}$. Note that this setup allows the densities (f_θ, h_ϕ) to be either probability density functions (pdf) or probability mass functions (pmf).

Dominated Likelihood Approximation

Definition. In the setup above, if Assumptions 1 and 2 below are satisfied, we say that the model H_ϕ is a dominated likelihood approximation (DLA) for the model F_θ.

Assumption 1: Broadened Support	$\text{supp}(H_\phi) \supsetneq \text{supp}(F_\theta), \quad \forall \phi \in \Phi, \theta \in \Theta.$
Assumption 2: Dominated Likelihood	$h_\phi(y) \leq f_\phi(y) \quad \forall\, y \in \mathcal{Y}, \phi \in \Phi.$

Intentional model misspecification. We can often safely assume that the target model has *well-specified support*

$$\text{supp}(F_\theta) = \text{supp}(G), \quad \forall \theta \in \Theta$$

in which case we have introduced an *intentional model misspecification*; we intentionally do inference with a model that has inflated support.

Maximum Likelihood Inference

Result.

We can substitute the maximum likelihood parameters for H into F.

This strategy provably minimizes an upper bound on an error term between the true data generating distribution G and the now tractable model.

Justification. We have

$$\begin{aligned}
& h_\phi(y) \leq f_\phi(y) \qquad \forall\, y \in \mathcal{Y}, \phi \in \Phi && \text{Assumption 2} \\
\implies\ & \mathbb{E}_G[\log h_\phi(Y)] < \mathbb{E}_G[\log f_\phi(Y)] && \text{Monotonicity, Assumption 1} \\
\iff\ & \mathbb{E}_G[\log g(Y)] - \mathbb{E}_G[\log f(Y)] < \mathbb{E}_G[\log g(Y)] - \mathbb{E}[\log h_\phi(Y)] && \text{algebra} \\
\iff\ & \texttt{KL}(G \,||\, F_\phi) < \texttt{KL}(G \,||\, H_\phi) && \text{def. KL}
\end{aligned}$$

where for simplicity we have assumed that G has density g.

Now by classic statistical results, the quasi-maximum likelihood estimator (QMLE)

$$\widehat{\phi}_n \triangleq \underset{\phi \in \Phi}{\text{argmax}} \frac{\sum_{i=1}^n \log h_\phi(Y_i)}{n}$$

asymptotically minimizes $\texttt{KL}(G \,||\, H_\phi)$. 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 $\texttt{KL}(G \,||\, F_\phi)$.

Bayesian Inference

Result

The posterior under likelihood H approximates the posterior under likelihood F

by maximizing a lower bound on the marginal likelihood of the target model F.

Justification. Given a prior distribution π on Φ, we obtain the following marginal density relationship from Assumption 2:

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

where we have used the same prior π on both Φ and Θ, using the implication from Assumption 2 that $\Phi \subseteq \Theta$. Hence, for any probability distribution Q on Φ within some chosen family $\mathcal{Q}$, we have

$$\log p_F(y) \geq \log p_H(y) \geq \texttt{ELBO}_H(Q)$$

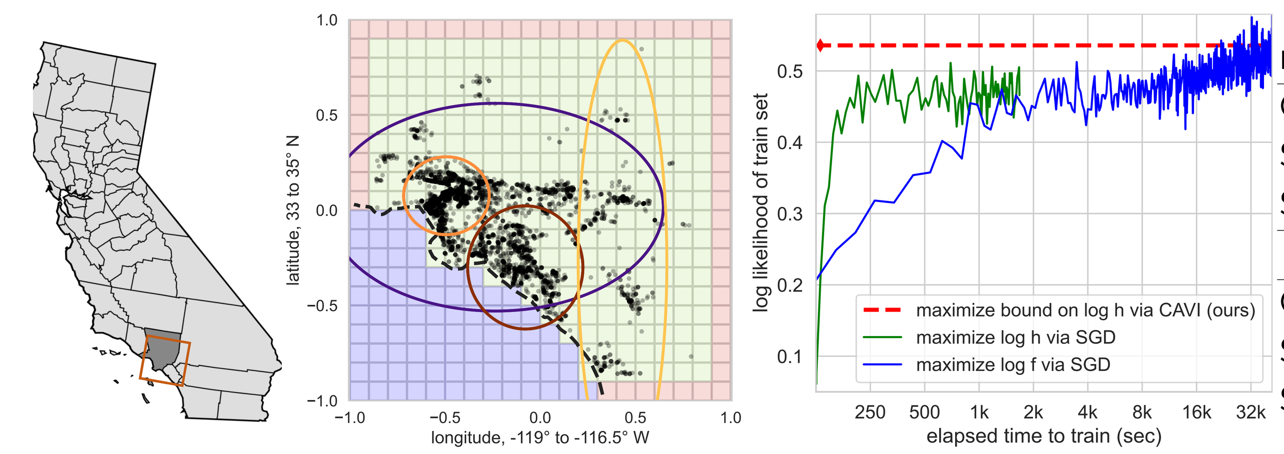
As is well-known, when the family $\mathcal{Q}$ is unconstrained, $\texttt{ELBO}_H(Q)$ is optimized by the true posterior under likelihood H. Thus, exact Bayesian inference for computing the posterior on Φ using H can be seen as producing the probability distribution on Φ which maximizes a lower bound on p_F.

Examples

We may start with fixed F_θ and produce H_ϕ (as in 1), or start with fixed H_ϕ and produce F_θ (as in 2).

	Target Model (F_θ)	**Approximation Model** (H_ϕ)
1	Truncated Gaussian	Gaussian
2	Categorical-from-binary GLMs	Independent Binary GLMs

Applications



method	train log f
CAVI on h	0.536
SGD on h	0.514
SGD on f	0.523
	test log f
CAVI on h	0.515
SGD on h	0.489
SGD on f	0.469

Figure 1. Application 1: Truncated mixtures of Gaussians for geolocations in southern CA. *Left:* Ideal model f truncates to the union of green rectangles (land area). Tractable model h (unconstrained mixture of Gaussians) allocates mass to water (blue) or out-of-bounds (red). Ellipses show the 99% high-density-areas of 4 Gaussian clusters fit to data using our CAVI approach. *Right:* Comparison of our approach to directly maximizing ideal likelihood f: DLA yields comparable models in far less time. Table reports each method's mean $\log f$ over all examples.

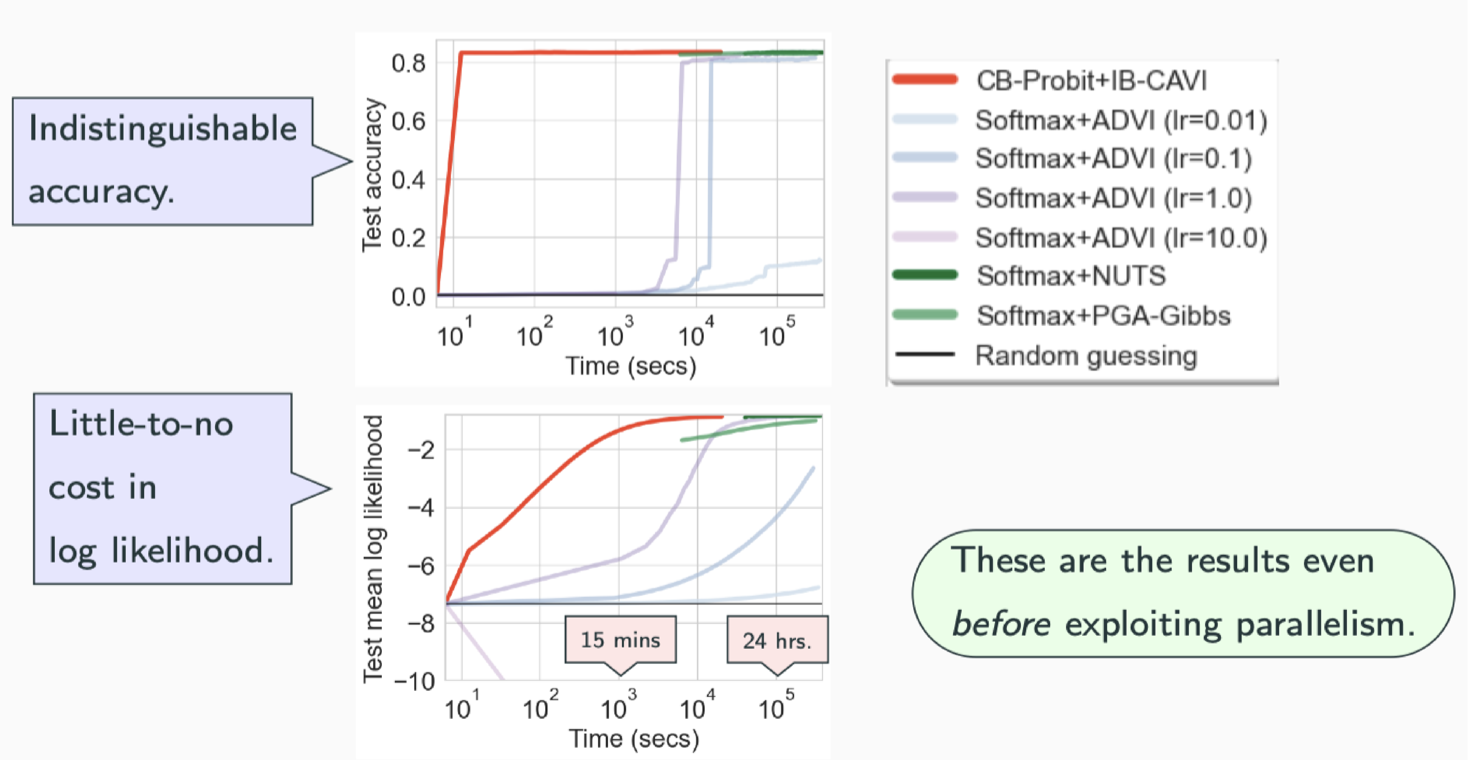


Figure 2. Application 2: Scalable Bayesian Categorical GLM for predicting computer process starts. Predicting a computer user's process starts (with 1,553 categories, 1,553 covariates, and 17,724 examples) in an cybersecurity intrusion application. We obtain quick inference by using CAVI on an independent binary model H, and substituting the posterior expectation into a newly defined categorical-from-binary (CB) model F which satisfies the DLA assumptions.

Michael.Wojnowicz@tufts.edu