PPL Bench

Evaluation Framework for Probabilistic Programming Languages

- Widespread adoption of statistical modeling in many domains
- Increasing number of PPLs, inference methods, models applied
- Need to quantify model performance across PPLs, inference methods

Conceptual Framework of PPL Bench:

- Model Instantiation and Data Generation

$$P_\theta(X, Z) = P_\theta(Z)P_\theta(X|Z)$$
$$Z_1 \sim P_\theta(Z)$$
$$X_{train} \overset{iid}{\sim} P_\theta(X|Z = Z_1)$$
$$X_{test} \overset{iid}{\sim} P_\theta(X|Z = Z_1)$$
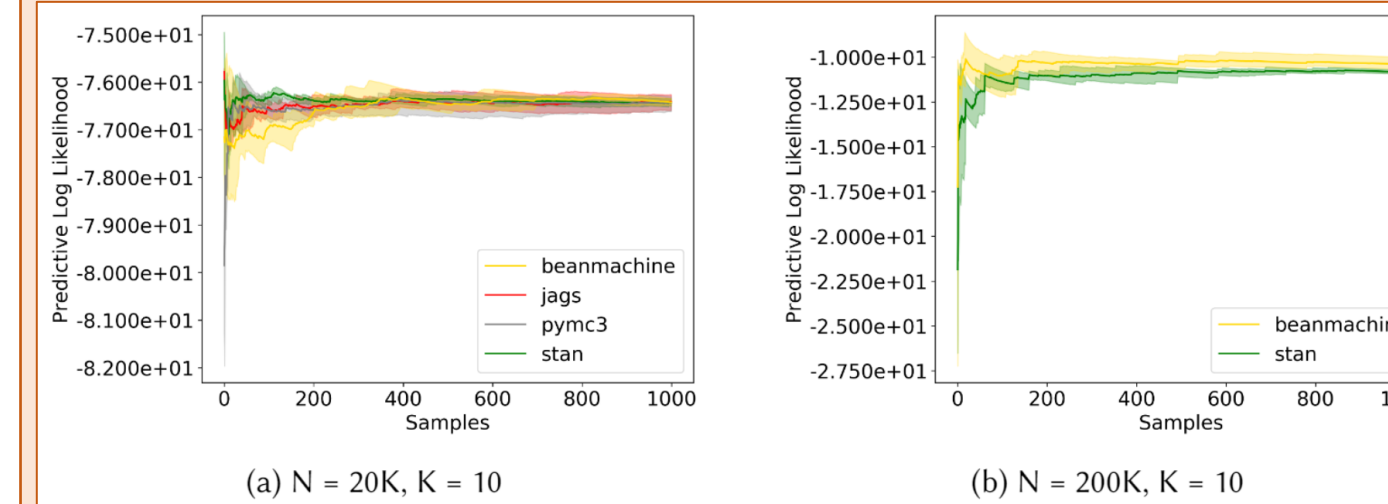
- PPL Implementation and Posterior Sampling

$$Z_{1...n}^* \sim P_\theta(Z|X = X_{train})$$

- Evaluation of Posterior Samples

$$\text{Predictive Log Likelihood}(n) = \log\left(\frac{1}{n}\sum_{i=1}^{n}(P(X_{test}|Z = Z_i^*))\right)$$
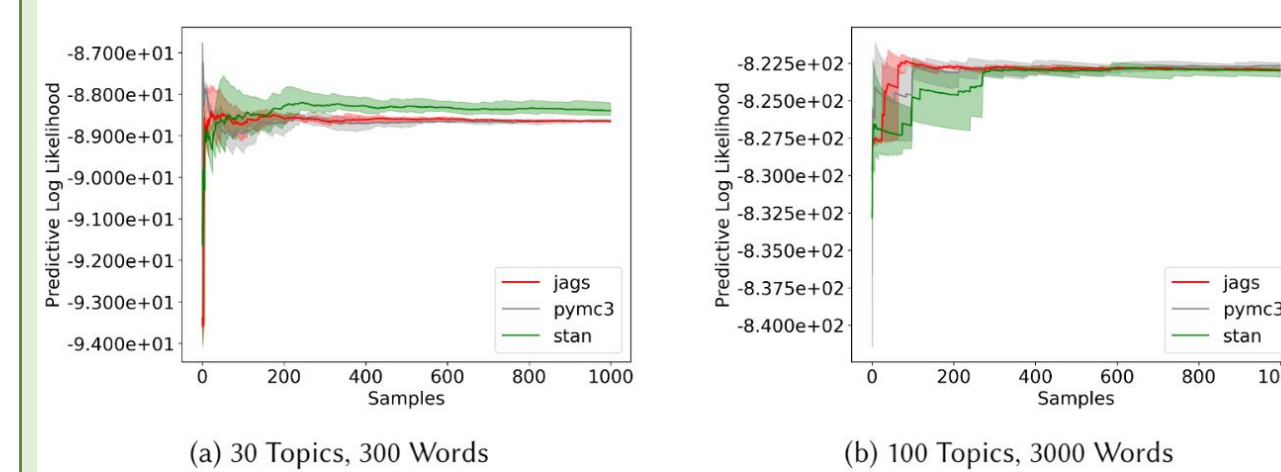
## Results and Insights

Bayesian Logistic Regression

(a) N = 20K, K = 10

(b) N = 200K, K = 10

PPL	N	Time(s)	n_{eff}/s		
			min	median	max
BeanMachine	20K	298.64	21.74	232.85	416.98
Stan	20K	47.06	21.62	25.78	108.63
Jags	20K	491.22	0.17	0.18	4.04
PyMC3	20K	48.37	23.65	27.10	67.27
BeanMachine	200K	1167.13	0.003	0.004	0.01
Stan	200K	10119.02	0.06	0.06	0.34

Noisy-Or Topic Model

(a) 30 Topics, 300 Words

(b) 100 Topics, 3000 Words

PPL	Words	Topics	Time(s)	n_{eff}/s		
				min	median	max
Stan	300	30	37.58	17.50	79.83	80.39
Jags	300	30	0.17	15028.96	17350.35	18032.36
PyMC3	300	30	38.05	39.50	78.83	79.33
Stan	3000	100	438.74	2.27	6.84	7.00
Jags	3000	100	0.68	3773.66	4403.28	4504.26
PyMC3	3000	100	298.79	3.96	10.04	14.72

Crowd-sourced Annotation Model
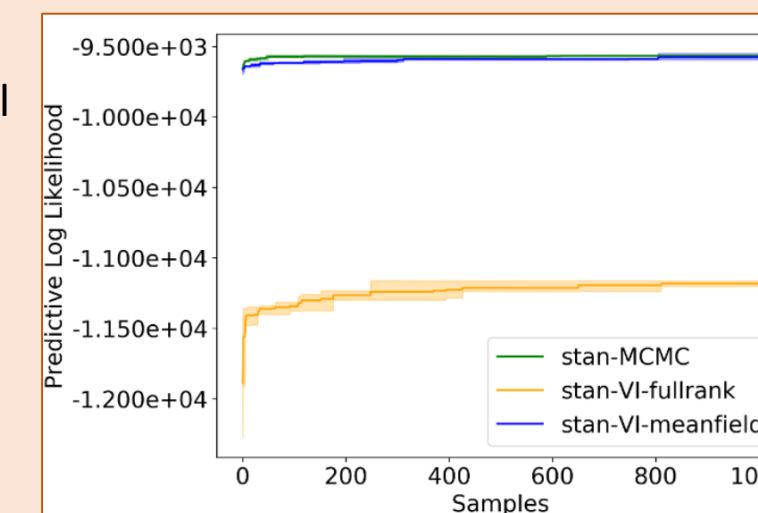
PPL	Time(s)	n_{eff}/s		
		min	median	max
Stan-MCMC	484.59	2.14	4.22	10.45
Stan-VI (meanfield)	15.23	0.48	7.86	216.35
Stan-VI (fullrank)	35.79	0.12	0.20	84.16

Design
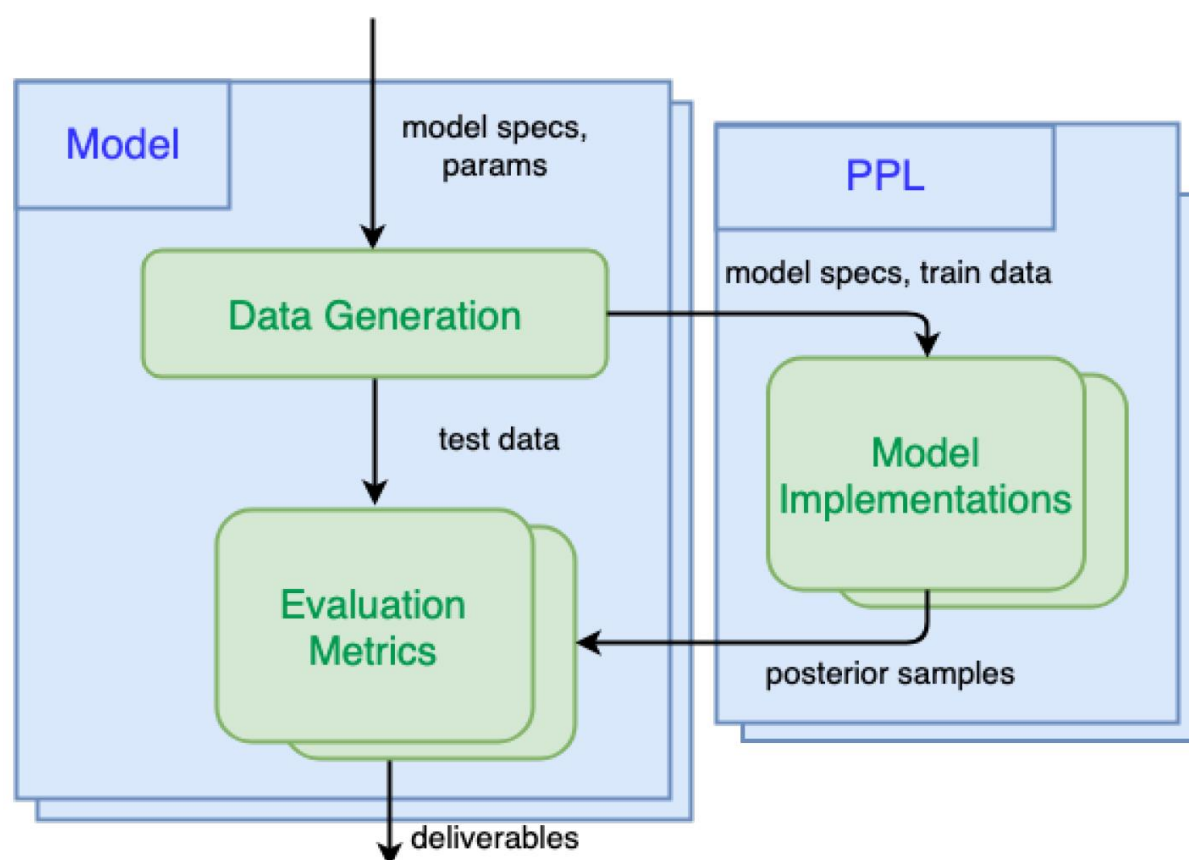
PPL Bench supports adding new PPLs
Currently compared:
Bean Machine, JAGS, PyMC3, STAN

Evaluation Metrics:

- Plot of the Predictive Log Likelihood w.r.t samples for each implemented PPL
- Gelman-Rubin convergence statistic r_{hat}
- Effective sample size n_{eff}
- Wall clock time per inference run

Models

PPL Bench supports adding new models
- Specify data generation
- Predictive Log Likelihood computation
- Other custom evaluation metrics

Bayesian Logistic Regression:
- Simple model; baseline
- Log-concave posterior, easy convergence

Noisy-Or Topic Model:
- Inferring topics from words in document
- Bayesian Network structure with topics and words as nodes
- Supports hierarchical topics

Crowdsourced Annotation:
- Inferring true label of an object given multiple labeler's label assignments
- Maintain confusion matrix of each labeler
- Includes inferring the unknown prevalence of labels

## Next Steps: PPL Bench Open Source

PPL Bench Summary:
- Modular Customizable framework for evaluating PPLs, statistical models, and inference methods
- Open source, with lot of potential for extension

Using PPL Bench:
- Comparing model performance across PPLs
- Comparing effectiveness of inference algorithms across models
- Evaluating new inference algorithms

Contributing to PPL Bench:
- Add new models
- Add model-specific eval metrics
- Add new PPLs
- Add PPL implementations of existing Models
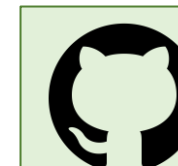
GitHub Repository:
github.com/facebookresearch/pplbench

Website:
pplbench.org

arXiv paper link:
https://arxiv.org/abs/2010.08886