# Diagnostic tools for approximate Bayes

Jason Clarkson, Bryan Liu and Mélodie Monod

Oxford University and Imperial College London

October 24, 2019

# Outline

# Introduction

Consider the usual objects in Bayesian inference:

- prior distribution $\phi \sim \pi(\cdot)$
- likelihood distribution observed data $y$ generated from $p(\cdot|\phi)$
- posterior distribution $\pi(\cdot|y)$

We present three diagnostics to evaluate the quality of the Bayes posterior approximation from either an MCMC or a VI.

1. Simulation-based Calibration
2. Pareto Smoothed Importance Sampling
3. Posterior Inference via Credible Intervals

# Introduction

We use as toy example a Bayesian linear regression, where the data are generated as:

$$\beta \sim \mathcal{N}(0, 10^2), \tag{1}$$
$$\alpha \sim \mathcal{N}(0, 10^2), \tag{2}$$
$$y \sim \mathcal{N}(\alpha + \beta x, 1.2^2), \tag{3}$$

where $x$ is some specified covariate.

# Simulation-based Calibration

Say we have the following:

- Exact prior: $\phi \sim \pi(\cdot)$
- Simulated data conditioned on prior: $y \sim p(\cdot|\phi)$
- Estimated posterior conditioned on data: $\theta \sim \tilde{\pi}(\cdot|y)$

Cook et al. (2006) said if the posterior is exact, integrating it over all possible prior and likelihood will return the prior. (i.e. data-averaged posterior $\equiv$ prior)

Talts et al. (2018) said if the above is true, the rank of one prior sample amongst many posterior samples will be uniformally distributed.

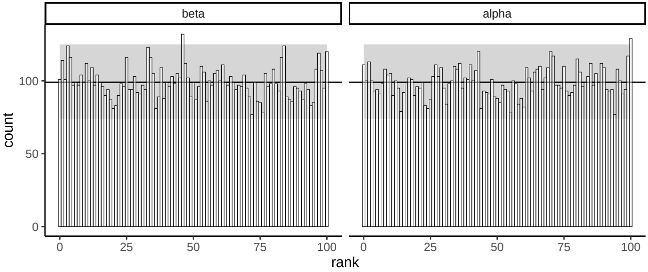
# Result is reproducible when everything is right



For 1:$N$ —

1. Sample a ground truth from prior: $\phi \sim \pi(\cdot)$
2. Simulate data using $\phi$: $y_{1:s} \sim p(\cdot|\phi)$
3. Obtain many posterior samples after fitting model with data: $\theta_1, \cdots, \theta_L \sim \tilde{\pi}(\cdot|y_{1:s})$
4. Compute rank of $\phi$ relative to $\{\theta_1, \cdots, \theta_L\}$

Check if the distribution of ranks is uniform.

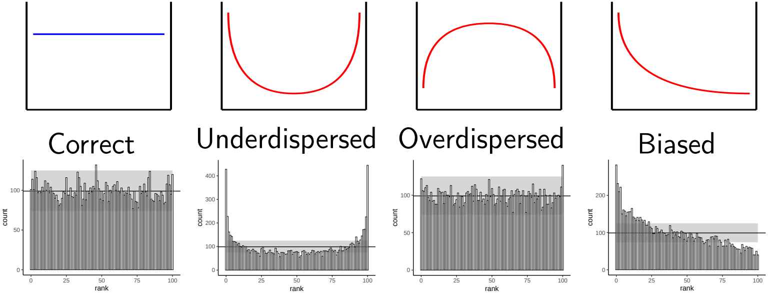


# Maybe not for a overdispersed prior...

Using a Bayesian linear regression example:

- Generate: $\beta, \alpha \sim \mathcal{N}(0, 10^2)$, $y_{1:s} \sim \mathcal{N}(X\beta + \alpha, 1.2^2)$
- Fit: $\beta \sim$ <whoops>, $\alpha \sim \mathcal{N}(0, 10^2)$ as prior, find posterior

Expected (top) & actual (bottom) rank distributions:

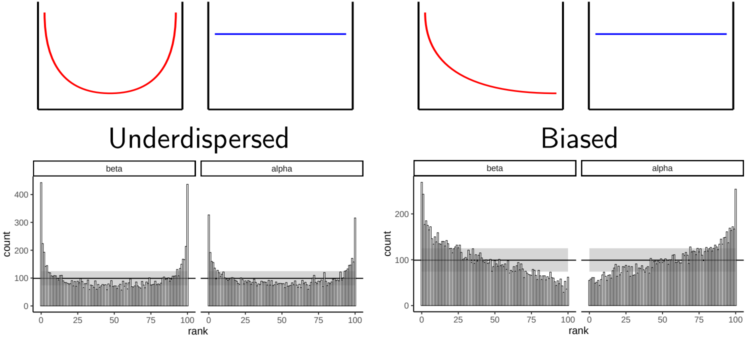


# Can SBC spot which parameter is wrong?

Using a Bayesian linear regression example:

- Generate: $\beta, \alpha \sim \mathcal{N}(0, 10^2)$, $y_{1:s} \sim \mathcal{N}(X\beta + \alpha, 1.2^2)$
- Fit: $\beta \sim$ <whoops>, $\alpha \sim \mathcal{N}(0, 10^2)$ as prior, find posterior

Expected (top) & actual (bottom) rank distributions:



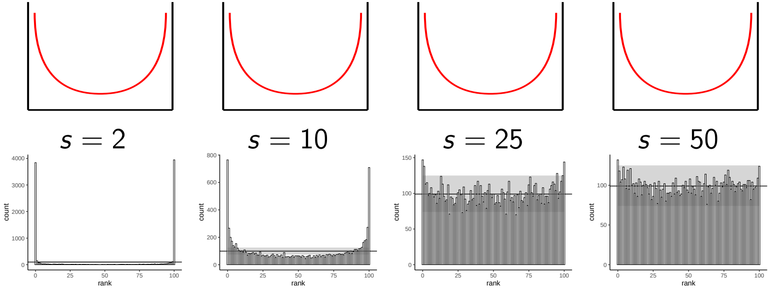
# Could a misspecified prior be masked by an abundance of data?



Using a Bayesian linear regression example:

- Generate: $\beta, \alpha \sim \mathcal{N}(0, 10^2)$, $y_{1:s} \sim \mathcal{N}(X\beta + \alpha, 1.2^2)$
- Fit: $\beta \sim$ <whoops>, $\alpha \sim \mathcal{N}(0, 10^2)$ as prior, find posterior

Expected (top) & actual (bottom) rank distributions:



# Variational Inference

## Definition

Variational inference (VI) consider a family of simple densities and find the member closest to the posterior.

Consider density $q(\theta|\psi)$, parametrized by $\psi$.
The Kullback-Leibler (KL) divergence with the posterior is,

$$\mathsf{KL}\big(q(\theta|\psi)||\pi(\theta|y)\big) = E_q\left[\log \frac{q(\theta|\psi)}{\pi(\theta|y)}\right]. \tag{4}$$

The evidence lower bound (ELBO) is,

$$\mathsf{ELBO}(\psi) = E_q\Big[\log p(y|\theta)\Big] - \mathsf{KL}\big(q(\theta|\psi)||\pi(\theta)\big). \tag{5}$$

By maximizing the ELBO, we encourage the optimization process to choose a candidate distribution which

1. explains the observed data well
2. is similar to the prior distribution

# Importance weights

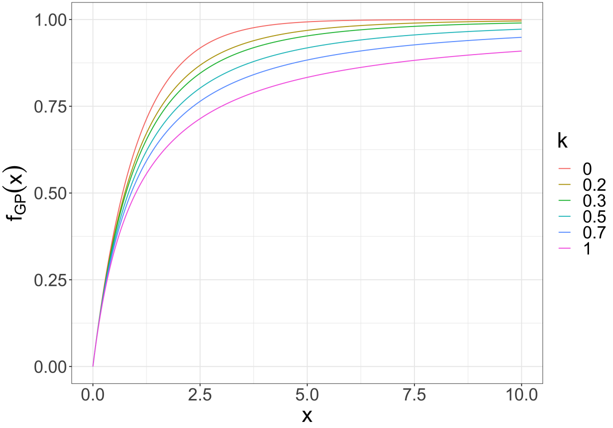
$$w_s = \frac{p(\theta_s, y)}{q(\theta_s|\psi^*)} \tag{6}$$

for $\theta_s \in \{\theta_1, ..., \theta_S\}$ evaluation of the proposal.

- They capture 'how close' the proposal is from the target.
- The success of plain importance sampling depends entirely on how many moments the importance ratios $r_s$ possess.
- The existence of the moments is by no means guaranteed.

# Generalized Pareto distribution fit

We fit $w_s | w_s > u$ with a Generalized Pareto of density.
The Generalized Pareto distribution possesses $\lfloor 1/k \rfloor$ finite moments when $k > 0$.



# Diagnostics for divergence



Then, $\hat{k}$ approximates

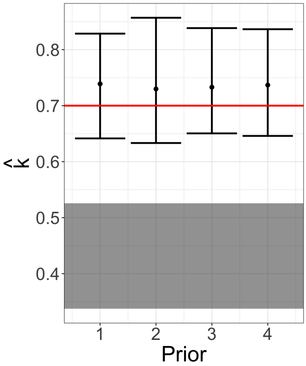
$$k = \inf\left\{ k' > 0 : E_q\left[\left(\frac{p(\theta|y)}{q(\theta|\psi^*)}\right)^{1/k'}\right] < \infty \right\}, \qquad (7)$$

a measure of divergence between $p(\theta|y)$ and $q(\theta|\psi^*)$. Vehtari et al. (2017) and Yao et al. (2018) have defined the following thresholds:

- $\hat{k} < 0.5$: The VI posterior approximation is close enough to the true posterior.
- $0.5 < \hat{k} < 0.7$: The VI posterior approximation is not perfect but can be helpful.
- $0.7 > \hat{k}$: The VI posterior approximation is not good.

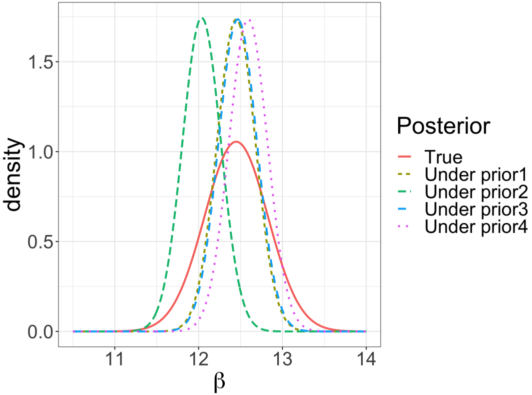
# Diagnostics for VI approximation under a Bayesian linear regression framework






Four prior specifications for $\beta$: (1) $\beta \sim \mathcal{N}(0, 10)$, (2) $\beta \sim \mathcal{N}(0, 2)$, (3) $\beta \sim \mathcal{N}(0, 100)$, (4) $\beta \sim \mathcal{N}(100, 10)$.

# Diagnostics for VI approximation under a Bayesian linear regression framework




Four prior specifications for $\beta$: (1) $\beta \sim \mathcal{N}(0, 10)$, (2) $\beta \sim \mathcal{N}(0, 2)$, (3) $\beta \sim \mathcal{N}(0, 100)$, (4) $\beta \sim \mathcal{N}(100, 10)$.

# Posterior Inference via Credible Intervals

Goal: Estimate CI using our posterior approximation $\tilde{\pi}(\cdot|y)$