Approximation Techniques for Bayesian Logistic Regression

Mauro Camara Escudero

Supervisor: Jonathan J. Forster

School of Mathematics, University of Southampton

Table of contents

1. Motivation and Problem Formulation
2. Stochastic Approximations
3. Deterministic Approximations
4. Results

Motivation and Problem Formulation

The Integration Problem

Many tasks in Bayesian Statistics can be seen as performing expectation with respect to a posterior distribution

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}\left[f(\boldsymbol{\theta})\right] = \int_{\Theta} f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$$

The resulting integration is often computationally intractable.

Stochastic Approximations

- *Monte Carlo*: Averages independent samples drawn from desired distribution.
- *Markov Chain Monte Carlo*: Monte Carlo on dependent samples produced by suitable Markov Chain.

Overview of Approximate Inference

Stochastic Approximations

- *Monte Carlo*: Averages independent samples drawn from desired distribution.
- *Markov Chain Monte Carlo*: Monte Carlo on dependent samples produced by suitable Markov Chain.

Deterministic Approximations

- *Laplace Approximation*: Approximates $p(\boldsymbol{\theta} \mid \mathrm{x})$ with normal distribution centered at the mode.
- *Variational Inference*: Approximates $p(\boldsymbol{\theta} \mid \mathrm{x})$ with the closest distribution $q(\boldsymbol{\theta})$ according to some objective function.

Bayesian Logistic Regression

Relationship between explanatory variables and response modelled with a Generalized Linear Model:

- Exponential Family distribution: $Y_i \sim \text{Bernoulli}(\pi_i)$
- Link Function: Logit $g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathsf{x}_i^\top \boldsymbol{\beta}$

Choose a normal prior $\boldsymbol{\beta} \sim \mathcal{N}(\mu_0, \Sigma_0)$.

Stochastic Approximations

- Pseudo-Random number generators allow us to sample from a uniform distribution $\mathcal{U}(0, 1)$.

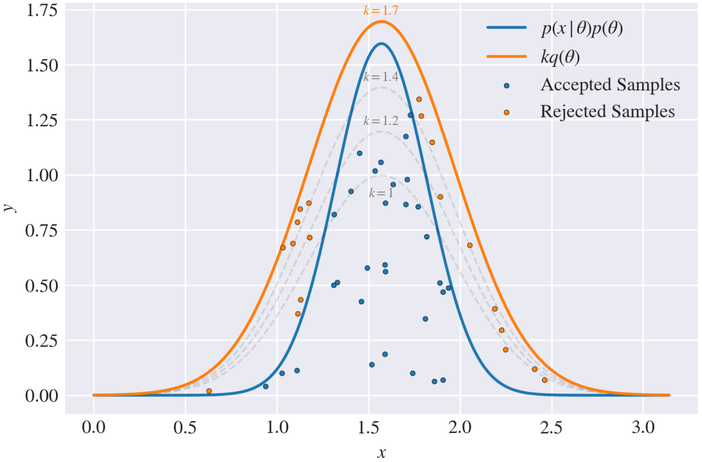
Sampling Methods

- Pseudo-Random number generators allow us to sample from a uniform distribution $\mathcal{U}(0, 1)$.
- Samples from non-uniform distributions can be obtained by transforming uniform samples but not always feasible, and computationally demanding.

Sampling Methods

- Pseudo-Random number generators allow us to sample from a uniform distribution $\mathcal{U}(0, 1)$.
- Samples from non-uniform distributions can be obtained by transforming uniform samples but not always feasible, and computationally demanding.
- Rejection Sampling draws uniform samples from a *function* whose graph is always above that of $p(\boldsymbol{\theta} \mid \mathrm{x})$ and accepts them as samples from $p(\boldsymbol{\theta} \mid \mathrm{x})$ if they are also under its graph.

Rejection Sampling Example



Monte Carlo Methods

- MONTE CARLO method uses $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ independent samples from $p(\boldsymbol{\theta} \mid \mathrm{x})$ to approximate the expectation

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathrm{x})}\left[f(\boldsymbol{\theta})\right] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i)$$

  this approximation converges by the law of large numbers.

Monte Carlo Methods

- MONTE CARLO method uses $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ independent samples from $p(\boldsymbol{\theta} \mid \mathrm{x})$ to approximate the expectation

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathrm{x})}\left[f(\boldsymbol{\theta})\right] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i)$$

  this approximation converges by the law of large numbers.
- UNIFORM SAMPLING: Rather than $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta} \mid \mathrm{x})$, draw $\boldsymbol{\theta}_i$ uniformly in Θ and use a weighted average of $f(\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i \mid \mathrm{x})$ instead. Inefficient because most time spent in regions of Θ with negligible mass.

Monte Carlo Methods

- Monte Carlo method uses $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ independent samples from $p(\boldsymbol{\theta} \mid \mathrm{x})$ to approximate the expectation

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathrm{x})}\left[f(\boldsymbol{\theta})\right] \approx \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i)$$

  this approximation converges by the law of large numbers.
- Uniform Sampling: Rather than $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta} \mid \mathrm{x})$, draw $\boldsymbol{\theta}_i$ uniformly in Θ and use a weighted average of $f(\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i \mid \mathrm{x})$ instead. Inefficient because most time spent in regions of Θ with negligible mass.
- Importance Sampling: Draw samples from a proposal distribution $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta})$ resembling $p(\boldsymbol{\theta} \mid \mathrm{x})$ and uses a weighted average with whose weights w_i compensate the error introduced by sampling from wrong distribution. Still inefficient in high dimensions.

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC) trades off the costly independence and exactness of the samples for approximate, dependent samples that are cheaper to compute.

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC) trades off the costly independence and exactness of the samples for approximate, dependent samples that are cheaper to compute.
- $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ generated by a Markov Chain whose equilibrium distribution is $p(\boldsymbol{\theta} \mid \mathrm{x})$.

Markov Chain Monte Carlo

- Markov Chain Monte Carlo (MCMC) trades off the costly independence and exactness of the samples for approximate, dependent samples that are cheaper to compute.
- $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ generated by a Markov Chain whose equilibrium distribution is $p(\boldsymbol{\theta} \mid \mathrm{x})$.
- Random Walk Metropolis-Hastings (RWMH) iteratively draws a sample from a symmetic proposal distribution $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}_i)$ depending only on the current sample value $\boldsymbol{\theta}_i$. Accepts $\boldsymbol{\theta}^*$ with a probability that makes the chain converge to $p(\boldsymbol{\theta} \mid \mathrm{x})$.

Deterministic Approximations

Laplace Approximation

Approximates $p(\boldsymbol{\theta} \mid \mathrm{x})$ with a multivariate normal distribution

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_0, -H(\boldsymbol{\theta}_0)^{-1})$$

centered at the mode $\boldsymbol{\theta}_0$ and with variance-covariance matrix

$$-\nabla^2 \ln(p(\mathrm{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}))^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

the negative inverse Hessian matrix of $\ln(p(\mathrm{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}))$.

Laplace Approximation for Bayesian Logisitic Regression

In Bayesian Logistic Regression the multivariate normal distribution is given by

$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, -H(\boldsymbol{\beta}_0)^{-1})$$

with

$$-H(\boldsymbol{\beta}_0) = \Sigma_0{}^{-1} + \sum_{i=1}^{n} \pi_i(\boldsymbol{\beta}_0)(1 - \pi_i(\boldsymbol{\beta}_0))\mathsf{x}_i \mathsf{x}_i^\top$$

Variational Methods

- Define a family of distributions $\mathcal{D}$.

Variational Methods

- Define a family of distributions $\mathcal{D}$.
- Define objective function measuring distance between $p(\boldsymbol{\theta} \mid \mathrm{x})$ and $q(\boldsymbol{\theta}) \in \mathcal{D}$. Usually KL divergence.

$$KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathrm{x})) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathrm{x})}\right)\right]$$

Variational Methods

- Define a family of distributions $\mathcal{D}$.
- Define objective function measuring distance between $p(\boldsymbol{\theta} \mid \mathrm{x})$ and $q(\boldsymbol{\theta}) \in \mathcal{D}$. Usually KL divergence.

$$KL(q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathrm{x})) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln\left(\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathrm{x})}\right)\right]$$

- Choose distribution minimizing KL divergence, or equivalently choose distribution *maximizing* Evidence Lower Bound (ELBO)

$$\mathrm{elbo}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta} \mid \mathrm{x})) = \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln\left(p(\mathrm{x} \mid \boldsymbol{\theta})\right] - \mathrm{KL}\left(q(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta})\right)\right.$$

Local Variational Methods for Bayesian Logistic Regression

- Bound each log success probability $\ln(\pi_i)$ in the likelihood with a quadratic function of $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ depending on variational parameters ξ_i.

Local Variational Methods for Bayesian Logistic Regression

- Bound each log success probability $\ln(\pi_i)$ in the likelihood with a quadratic function of $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ depending on variational parameters ξ_i.
- This results in a Gaussian bound for the posterior.

$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v)$$

with $\boldsymbol{\mu}_v$ and $\boldsymbol{\Sigma}_v$ depending on the ξ_i's.

Local Variational Methods for Bayesian Logistic Regression

- Bound each log success probability $\ln(\pi_i)$ in the likelihood with a quadratic function of $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ depending on variational parameters ξ_i.
- This results in a Gaussian bound for the posterior.

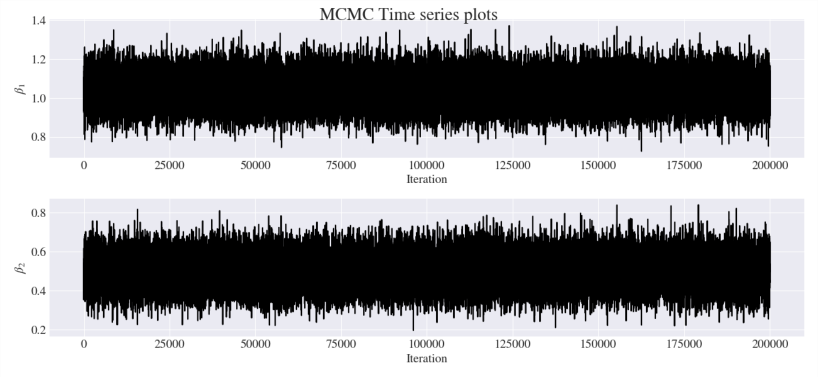
$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_v, \Sigma_v)$$

  with $\boldsymbol{\mu}_v$ and Σ_v depending on the ξ_i's.
- Maximize this bound using the EM algorithm

Results

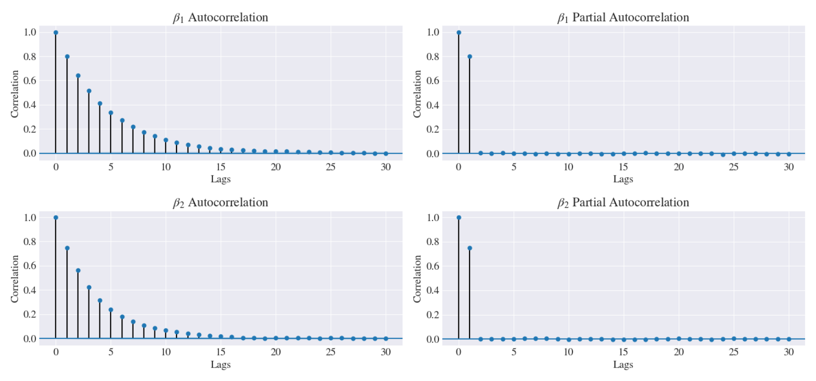
Models with Explanatory Variables

RWMH convergence diagnosed via trace plots.

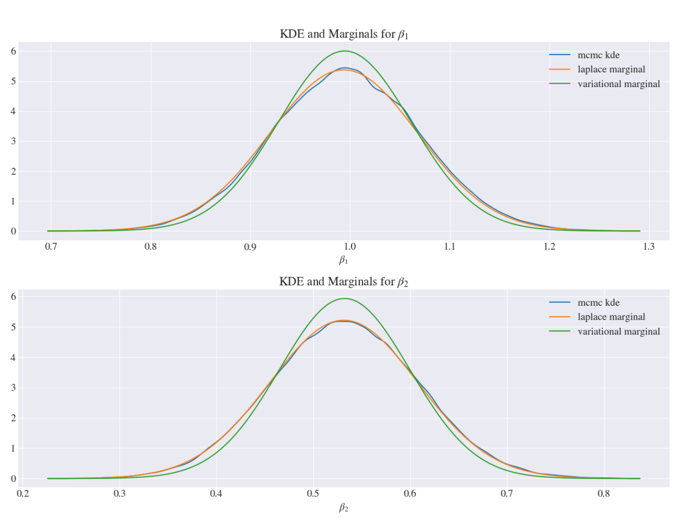


Models with Explanatory Variables

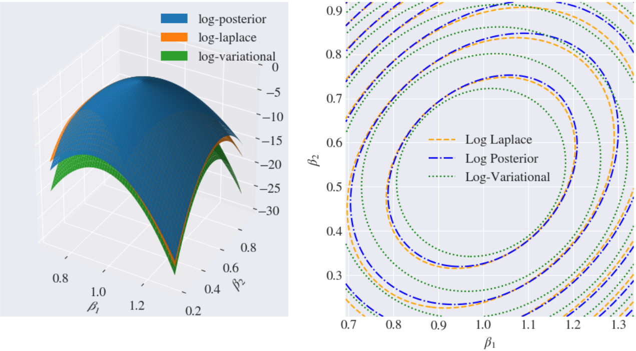
Markov Property diagnosed with Auto-Correlation and Partial Auto-Correlation plots.



Models with Explanatory Variables



Models with Explanatory Variables



Conclusion

Laplace outperforms Variational, which underestimates the tails.

Model without Explanatory Variables

- Posterior distribution

$$p(\beta \mid \mathrm{y}) = \frac{1}{B(n\bar{y}, n - n\bar{y})} \frac{e^{\beta n\bar{y}}}{\left(1 + e^{\beta}\right)^{n}}$$

Model without Explanatory Variables

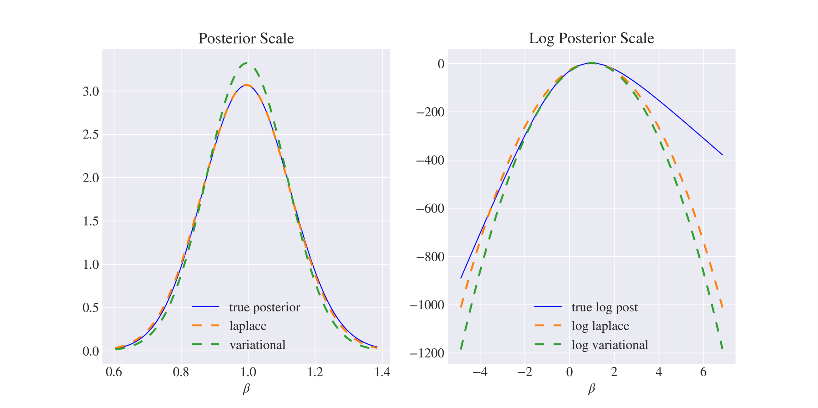
- Posterior distribution

$$p(\beta \mid \mathbf{y}) = \frac{1}{B(n\bar{y}, n - n\bar{y})} \frac{e^{\beta n \bar{y}}}{(1 + e^{\beta})^n}$$

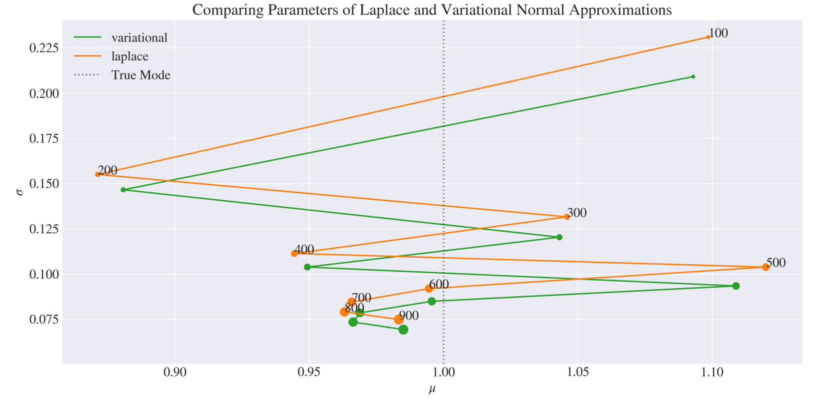
- Laplace approximation

$$q_l(\beta) = \mathcal{N}\left(\ln\left(\frac{\bar{y}}{1 - \bar{y}}\right), \frac{1}{n\bar{y}(1 - \bar{y})}\right)$$

Model without Explanatory Variables



Model without Explanatory Variables



Conclusion

Variational Mean closer to true population mode β, especially for small data set sizes.

Further Work

- Multimodal posterior distributions.

Further Work

- Multimodal posterior distributions.
- Different variational bound.

Further Work

- Multimodal posterior distributions.
- Different variational bound.
- Relationship between Laplace and Variational distributions.

Questions?

Backup Slide - Variational EM

Algorithm 1: Variational Approximation

Initialize $\boldsymbol{\xi} = (\xi_1^{(1)}, \ldots, \xi_n^{(1)})^\top \in \mathbb{R}^n$ randomly.

for $j = 1, 2, \ldots, \Delta$:

  for $i = 1, 2, \ldots, n$ **do**:

    Find mean and variance-covariance matrix and update variational parameters.

$$\Sigma_v = \left(2 \sum_{i=1}^{n} \lambda(\xi_i^{(j)}) \mathbf{x}_i \mathbf{x}_i^\top + \Sigma_0^{-1} \right)^{-1}$$

$$\boldsymbol{\mu}_v = \Sigma_v \left(\sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \frac{1}{2} \right) + \Sigma_0^{-1} \boldsymbol{\mu}_0 \right)$$

$$\xi_i^{(j+1)} = \sqrt{\mathbf{x}_i^\top \left(\Sigma_v + \boldsymbol{\mu}_v \boldsymbol{\mu}_v^\top \right) \mathbf{x}_i}$$

  end

end