Online Bayesian Variable Selection and Bayesian Model Averaging for Streaming Data

Joyee Ghosh [*], Aixin Tan [†]and Lan Luo [‡]

Abstract

There is an increasing prevalence of streaming data generation in diverse fields like healthcare, finance, social media, and weather forecasting. In order to acquire helpful insights from these massive datasets, timely analysis is essential. In this article, we assume that the streaming data are analyzed in batches. Traditional offline methods, which involve storing and analyzing all individual records, can be repeatedly applied to the cumulative data, but encounter significant challenges in storage and computing costs. Existing online methods offer faster approximations but most methods neglect model uncertainty, causing overconfidence and instability. To bridge this gap, we propose novel online Bayesian approaches that incorporate model uncertainty within a Bayesian model averaging (BMA) framework, for generalized linear models (GLMs). We propose computationally efficient methods to update the posterior, with individual records from the latest batch of data and summary statistics from previous batches. We demonstrate using simulation studies and real data that our methods can offer much faster analysis compared to traditional methods, with no substantial drop in accuracy.

Key words: Generalized linear models, Logistic regression, Posterior inclusion probability, Markov Chain Monte Carlo, Screening, Separation.

1 Introduction

Streaming data refers to data that are being generated continuously over time. For example, data generated by smart devices such as apple watch or fitbit would fall under this category. We assume scenarios in which data are collected in batches over time, and the primary

[*] Associate Professor, The University of Iowa

[†] Associate Professor, The University of Iowa

[‡] Assistant Professor, Rutgers University

Joyee Ghosh and Aixin Tan have equal contributions.

goal is to analyze the full dataset after each new batch of data arrives. If data on all individual records are saved for all batches, conventional methods can be used for statistical analysis, which is also referred to as offline methods. For long streams of data, storing all individual records may require infeasibly costly amounts of disk space. Furthermore, repeatedly analyzing such large datasets using conventional methods can be computationally intensive and so slow that the method is no longer useful. Online methods treat conventional methods as the gold standard in terms of accuracy, and offer faster and computationally more feasible approximations to this benchmark. Most existing online methods analyze the data using a single model and ignore model uncertainty, which leads to issues like overconfidence, potentially misleading inference and model instability. Thus there is a critical need for development of online methods that take into account model uncertainty, which can be naturally incorporated into the prior probability distribution in a BMA framework. Relying solely on analyzing the entire data stream with a single model could result in researchers and decision-makers losing or distorting crucial data characteristics.

In this article, our goal is to develop computationally efficient methods to analyze streaming data with Bayesian GLMs, specifically addressing variable selection uncertainty. We propose online Bayesian methods that update the posterior distribution over the model space, using individual records from the most recent batch of data and summary statistics from past batches. Our method would enable data analysts to employ more flexible models for streaming data without the computational burdens associated with conventional methods.

Online learning methods are designed to analyze data that arrive in sequential batches to answer successive questions of interest while mitigating storage and computing cost. In the realm of parametric models, many online learning methods focus on renewed estimation of unknown parameters using the maximum likelihood principle or its regularized versions, based on the newest batch of data and summary statistics of the historical data.

In the context of GLMs, some examples of online learning solutions include the stochastic gradient descent (SGD) (Robbins and Monro, 1951); an enhanced version called the implicit SGD (Toulis et al., 2014) which is more robust to learning rate mis-specification; online iterative algorithms for linear models and estimating equations (Schifano et al., 2016), an online one step iteratively reweighted least squares (IRWLS) (Zhang and Yang, 2021), and a renewable estimator (Luo and Song, 2020) that enjoys smaller bias and standard errors, and provides richer inference like confidence intervals.

Several methods have been proposed for online model selection. These include online criterion based (such as AIC, BIC, and DIC) variable selection methods for linear regression models (Wang et al., 2016), recursive model selection for high dimensional GLMs with estimation of parameters via score equations (Shi et al., 2021), and more recently the debiased stochastic gradient descent and the online debiased lasso (Luo et al., 2023; Han et al., 2024), which conducts variable selection and inference for high dimensional GLMs, using only summary statistics of the historical data. The majority of these methods perform non Bayesian variable selection, with the exception of DIC (Wang et al., 2016), but the latter has been developed in the context of linear regression models.

The Bayesian approach has advantages in addressing uncertainty more explicitly via prior distributions. Despite substantial literature in BMA, which we introduce in Section 2, there is limited previous work on online BMA methods. One of the earliest papers on online model selection is by Sato (2001), who combined the idea of sequential model selection with online variational Bayes, and proposed an online model selection algorithm, which was applied to a mixture of Gaussian models. McCormick et al. (2012) considered dynamically weighted model averages for logistic regression, where the parameters for each model are treated as latent variables that follow their own state space models. Onorante and Raftery (2016) proposed a dynamic version of Occam's window for dynamic linear models, when the original version of dynamic model averaging becomes computationally intensive with many predictors. In the dynamic model averaging setting, the data generating mechanism changes at every time point. In contrast, we focus on models that have parameter values that are stable over a long term, and use the idea of renewable estimation for online updating of the posterior distribution over models.

In Section 2, we review the Bayesian approach to model uncertainty. In Section 3, we introduce our online methods for BMA in GLMs. Simulation studies for logistic regression are presented in Section 4, and an online analysis of occupant level data on traffic crashes is conducted in Section 5. Finally, a summary of the contributions in this paper and future directions are presented in Section 6.

2 Bayesian Variable Selection and Bayesian Model Averaging

In high dimensional model selection with many covariates, usually no single model stands out. A possible solution is to adopt a framework of BMA. This approach incorporates model uncertainty using a hierarchical probabilistic framework. Instead of selecting a single model, this framework allows us to combine models by taking a mixture of all models in the model space with posterior probabilities of models serving as the mixture weights, given the observed data. This has been empirically demonstrated to improve upon the predictive performance of a single model (Hoeting et al., 1999).

We provide a brief overview of the Bayesian approach to model uncertainty. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)'$ denote a vector of indicator variables, such that $\gamma_j = 1$ when the model includes the covariate $\boldsymbol{x}_j$ as a column in the design matrix $X_{\boldsymbol{\gamma}}$, and $\gamma_j = 0$ otherwise. Thus the entire list of 2^p models can be represented by all possible configurations of $\boldsymbol{\gamma}$, and is denoted by Γ. Consider a model $\boldsymbol{\gamma}$ with $q_{\boldsymbol{\gamma}}$ predictors for the response variable y, that is $\sum_{j=1}^{p} \gamma_j = q_{\boldsymbol{\gamma}}$, and let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denote the corresponding $q_{\boldsymbol{\gamma}}$ dimensional vector of nonzero regression coefficients. Given the data distribution for the response variable y, a prior distribution is specified on all unknown parameters: $\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$, $\boldsymbol{\gamma} \sim p(\boldsymbol{\gamma})$. Posterior inference is then based on $\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}|D)$, which is composed of two parts. Within each model $\boldsymbol{\gamma}$, the posterior on the model parameters is given as

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, D) = \frac{L(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; D)\, p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})}{\int L(\boldsymbol{\beta}'_{\boldsymbol{\gamma}}; D)\, p(\boldsymbol{\beta}'_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\, d\boldsymbol{\beta}'_{\boldsymbol{\gamma}}} = \frac{L(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; D)\, p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})}{m(D|\boldsymbol{\gamma})},$$

where $m(D|\boldsymbol{\gamma}) = \int L(\boldsymbol{\beta}'_{\boldsymbol{\gamma}}; D)\, p(\boldsymbol{\beta}'_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\, p(\boldsymbol{\gamma})\, d\boldsymbol{\beta}'_{\boldsymbol{\gamma}}$ is the marginal distribution of the data under model $\boldsymbol{\gamma}$, which is also referred to as the marginal likelihood corresponding to the model $\boldsymbol{\gamma}$.

Equipped with the marginal likelihood, the posterior distribution over models is given as

$$\pi(\boldsymbol{\gamma}|D) = \frac{m(D|\boldsymbol{\gamma})\, p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \boldsymbol{\Gamma}} m(D|\boldsymbol{\gamma}')\, p(\boldsymbol{\gamma}')}. \tag{1}$$

Even for static data sets with a moderate sample size n, evaluation of the denominators in the above equations poses computational challenges. **The first difficulty** is the lack of an analytical expression for the integral $m(D|\boldsymbol{\gamma})$ except in normal linear models with conjugate priors. Various approximations are available for $m(D|\boldsymbol{\gamma})$, see Raftery (1996) for more details. A commonly used approximation is provided by

$$\tilde{\pi}(\boldsymbol{\gamma}|D) = \frac{\exp\{-\frac{1}{2}\mathrm{BIC}(\boldsymbol{\gamma})\}p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \boldsymbol{\Gamma}} \exp\{-\frac{1}{2}\mathrm{BIC}(\boldsymbol{\gamma}')\}p(\boldsymbol{\gamma}')}, \tag{2}$$

where BIC($\boldsymbol{\gamma}$) is the usual Bayesian information criterion for model $\boldsymbol{\gamma}$, and given by

$$\mathrm{BIC}(\boldsymbol{\gamma}) = -2l(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}; D) + q_{\boldsymbol{\gamma}} \log(n)\,, \tag{3}$$

where $l(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}; D)$ is the loglikelihood under model $\boldsymbol{\gamma}$, evaluated at the MLE. This approximation allows us to calculate the relative posterior probability

$$\frac{\tilde{\pi}(\boldsymbol{\gamma}|D)}{\tilde{\pi}(\boldsymbol{\gamma}'|D)} = \exp\{-\frac{1}{2}(\mathrm{BIC}(\boldsymbol{\gamma}) - \mathrm{BIC}(\boldsymbol{\gamma}'))\}\frac{p(\boldsymbol{\gamma})}{p(\boldsymbol{\gamma}')}, \tag{4}$$

between any two models, $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$, but not the actual probabilities.

Obtaining the actual probabilities in (2) is **the second difficulty**. Analytically, this would require summing over all models in $\boldsymbol{\Gamma}$, in the denominator of (2). This calculation is possible but potentially expensive for $p < 25$, and prohibitively expensive for a larger value of p. For standard computers one would typically run out of computer memory for $p > 30$. A solution is to replace the enumeration of all 2^p terms by a stochastic exploration, e.g., a Markov chain that visits the $\boldsymbol{\gamma}$s with relatively high posterior probability, such that the invariant distribution of the chain is $\tilde{\pi}(\boldsymbol{\gamma}|D)$. In practice, the proportion of times the Markov chain visits a specific model $\boldsymbol{\gamma}$ can be used as a Monte Carlo estimate of $\tilde{\pi}(\boldsymbol{\gamma}|D)$. Note that the construction of such a Markov chain only requires the relative posterior probabilities in (4). For large p, MCMC (or other) algorithms have been adopted by several authors (George and McCulloch, 1993; Madigan and York, 1995; George and McCulloch, 1997; Raftery et al., 1997; Hans et al., 2007; Carvalho et al., 2010; Liu et al., 2014; Williams et al., 2023; Nie and Ročková, 2023) to explore promising regions of the model space. In the next section, we discuss how to **solve the above two difficulties in an online setting**.

3 Online BMA for Generalized Linear Models

Let y_i denote the response variable, and let $\boldsymbol{x}_i$ denote the covariates, for $i = 1, 2, \dots, n$. A generalized linear model (GLM) (Jørgensen, 1987; Agresti, 2013) assumes that dataset D consists of independent observations $(\boldsymbol{x}_i, y_i), i = 1, \dots, n$, such that conditional on $\boldsymbol{x}_i$, y_i are generated from

$$f(y_i|\theta_i, \psi) = \exp\left\{c(y_i, \psi) + \frac{y_i\theta_i - b(\theta_i)}{a(\psi)}\right\}, \quad i = 1, \ldots, n, \tag{5}$$

where the GLM links the mean $\mu_i = E(y_i|\theta_i, \phi)$ to a linear predictor ${\boldsymbol{x}_i}^T\boldsymbol{\beta}$ via a link function $g(.)$, such that $g(\mu_i) = {\boldsymbol{x}_i}^T\boldsymbol{\beta}$. For the canonical link function $g(\mu_i) = \theta_i$. Here $\boldsymbol{\beta}$ is the regression parameter of interest and ψ is called a dispersion parameter. Some components of $\boldsymbol{\beta}$ maybe zero, meaning that the corresponding covariates are not associated with the response variable. Some of our inference goals include estimating the regression parameters and predicting the response variables for new observations with observed covariates.

When ψ is known, such as in logistic or Poisson regression, the loglikelihood associated with dataset D is a summation of n terms, given by

$$l(\boldsymbol{\beta}; D) = \sum_{i=1}^{n} c(y_i, \psi) + \frac{1}{a(\psi)} \sum_{i=1}^{n} (y_i\theta_i - b(\theta_i))\,. \tag{6}$$

Let ∇ and $\nabla\nabla$ denote the gradient and the Hessian operator for differentiable multi-variable functions, respectively. Let $\hat{\boldsymbol{\beta}}$ denote the maximum likelihood estimator (MLE), such that $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}; D)$ and $\nabla l(\boldsymbol{\beta}; D) = \mathbf{0}$. Denote $J = -\nabla\nabla l(\boldsymbol{\beta}; D)$ and $\hat{J} = -\nabla\nabla l(\hat{\boldsymbol{\beta}}; D)$. For the remainder of this proposal we focus on models with a known dispersion parameter ψ, but the methods can be extended to the case of an unknown ψ by using a consistent estimator of ψ as in Luo and Song (2020).

Suppose at each time point $b = 1, \ldots, B$, a new batch of data D_b of size n_b arrives. The aggregated data by time b is $D_b^* = D_1 \cup \cdots \cup D_b$ of size $N_b = n_1 + \cdots + n_b$. For clarity, we will suppress the dependence of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ on $\boldsymbol{\gamma}$ in the notation in this subsection, and simply refer to it as $\boldsymbol{\beta}$. The loglikelihood associated with D_b is denoted by $l(\boldsymbol{\beta}; D_b)$, and that associated with D_b^* is $l(\boldsymbol{\beta}; D_b^*) = l(\boldsymbol{\beta}; D_{b-1}^*) + l(\boldsymbol{\beta}; D_b)$. Let $\hat{\boldsymbol{\beta}}_b = \arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}; D_b^*)$ denote the MLE at time b. Finally, at time point b, denote the observed Fisher information matrix and its value at the MLE by $J_b(D_b^*; \boldsymbol{\beta}) = -\nabla\nabla l(\boldsymbol{\beta}; D_b^*)$ and $\hat{J}_b = J_b(D_b^*; \hat{\boldsymbol{\beta}}_b)$ respectively.

3.1 Addressing the First Difficulty of Intractable Marginal Likelihood

Evaluation of the loglikelihood in (6) and its associated features like the gradient vectors and Hessian matrices is critical for inference in GLMs. However, the computation of these quantities becomes expensive when N_b is large as the loglikelihood by batch b involves a summation over N_b terms. To reduce the computational burden, one possible solution is to approximate quantities at time b based on subject level data only for the current batch of data D_b (of size n_b), together with summary statistics from historical data $D_{(b-1)}^*$. One of the main contributions of this work is to use such an online framework to perform approximate BMA. Based on (2) and (3), this task requires evaluating the following at each time b:

$$\hat{\boldsymbol{\beta}}_b \text{ (the MLE of } \boldsymbol{\beta} \text{ given the aggregate data } D_b^*\text{), and} \tag{7}$$

$$l(\hat{\boldsymbol{\beta}}_b; D_b^*)\,. \tag{8}$$

For the MLE in (7), Luo and Song (2020) introduced an online method called renewable estimation. The renewable estimate at time b, denoted by $\widetilde{\boldsymbol{\beta}}_b$, depends on the current data D_b and summary statistics of the historical data: an estimate of the regression coefficients, $\widetilde{\boldsymbol{\beta}}_{(b-1)}$, and the aggregated negative Hessian matrix evaluated at past estimates, $\tilde{J}_{(b-1)} = \sum_{j=1}^{b-1} J_j(D_j; \widetilde{\boldsymbol{\beta}}_j)$. Luo and Song (2020) showed that the renewable estimate is asymptotically equivalent to the MLE, as $N_b \to \infty$. Below, we provide two solutions to approximate the likelihood in (8) using quantities already available from the aforementioned renewable online method.

3.2 The Online 1 Method

At batch $b = 1$, we use the available dataset D_1 to calculate and record summaries $(\widetilde{\boldsymbol{\beta}}_1, l(\widetilde{\boldsymbol{\beta}}_1; D_1))$, where $\widetilde{\boldsymbol{\beta}}_1$ is the usual or offline MLE $\hat{\boldsymbol{\beta}}_1$ based on D_1. After this batch, the individual records in D_1 are deleted. At batch $b = 2$, we have access to the individual records in dataset D_2 and summaries from the previous batch. The loglikelihood evaluated at the renewable MLE, at batch $b = 2$, is given by

$$l(\widetilde{\boldsymbol{\beta}}_2; D_2^*) = l(\widetilde{\boldsymbol{\beta}}_2; D_1) + l(\widetilde{\boldsymbol{\beta}}_2; D_2)\,. \tag{9}$$

The first term $l(\widetilde{\boldsymbol{\beta}}_2; D_1)$ in the above equation cannot be calculated, because the individual records in the previous batch D_1 are no longer accessible. We propose to approximate it by a Taylor expansion:

$$l(\widetilde{\boldsymbol{\beta}}_2; D_1) \approx l(\widetilde{\boldsymbol{\beta}}_1; D_1) + (\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1)^T \nabla(\widetilde{\boldsymbol{\beta}}_1; D_1) = l(\widetilde{\boldsymbol{\beta}}_1; D_1) + 0\,,$$

where the second term is zero because the gradient of the loglikelihood evaluated at the MLE is $\mathbf{0}$. Hence our approximation for $l(\widetilde{\boldsymbol{\beta}}_2; D_2^*)$ at batch $b = 2$ is

$$\check{l}_2 = l(\widetilde{\boldsymbol{\beta}}_1; D_1) + l(\widetilde{\boldsymbol{\beta}}_2; D_2).$$

We record the summaries $(\widetilde{\boldsymbol{\beta}}_2, \check{l}_2)$ and discard the individual records in D_2. Moving on to time $b = 3$, we use the approximation

$$\begin{aligned} l(\widetilde{\boldsymbol{\beta}}_3; D_3^*) &= l(\widetilde{\boldsymbol{\beta}}_3; D_1) + l(\widetilde{\boldsymbol{\beta}}_3; D_2) + l(\widetilde{\boldsymbol{\beta}}_3; D_3) \\ &\approx l(\widetilde{\boldsymbol{\beta}}_1; D_1) + l(\widetilde{\boldsymbol{\beta}}_2; D_2) + l(\widetilde{\boldsymbol{\beta}}_3; D_3) = \check{l}_2 + l(\widetilde{\boldsymbol{\beta}}_3; D_3) =: \check{l}_3 \end{aligned}$$

and record $(\widetilde{\boldsymbol{\beta}}_3, \check{l}_3)$.

Extending the above technique to a generic batch b, where $b \geq 3$, we will have access to the observations in dataset D_b and summaries from the past batch $(b-1)$: $(\widetilde{\boldsymbol{\beta}}_{(b-1)}, \check{l}_{(b-1)})$. Here, $\check{l}_{(b-1)} = l(\widetilde{\boldsymbol{\beta}}_1; D_1) + l(\widetilde{\boldsymbol{\beta}}_2; D_2) + \cdots + l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)})$ serves as an approximation to $l(\widetilde{\boldsymbol{\beta}}_b; D_1) + l(\widetilde{\boldsymbol{\beta}}_b; D_2) + \cdots + l(\widetilde{\boldsymbol{\beta}}_b; D_{(b-1)})$. Hence

$$l(\widetilde{\boldsymbol{\beta}}_b; D_b^*) = l(\widetilde{\boldsymbol{\beta}}_b; D_1) + l(\widetilde{\boldsymbol{\beta}}_b; D_2) + \cdots + l(\widetilde{\boldsymbol{\beta}}_b; D_b) \approx \check{l}_{(b-1)} + l(\widetilde{\boldsymbol{\beta}}_b; D_b) = \check{l}_b \,.$$

We record the summaries $(\widetilde{\boldsymbol{\beta}}_b, \check{l}_b)$, which are the online estimates of the quantities in (7)-(8), and their values at batch b will be used for the inference for BMA after the bth batch of dataset arrives, for $b = 1, 2, \ldots, B$.

3.3 The Online 2 Method

To improve the accuracy of the first-order Taylor expansion of the loglikelihood $l(.)$ in the Online 1 approximation, the Online 2 method adopts a second-order expansion. The Online 2 method requires an additional Hessian matrix at each time b. Importantly, this Hessian matrix is already calculated and recorded in the online algorithm for the renewable estimator $\widetilde{\boldsymbol{\beta}}_b$, hence does not add much additional computing effort.

At batch $b = 1$, we use the available dataset D_1 to calculate and record the summaries $(\widetilde{\boldsymbol{\beta}}_1, l(\widetilde{\boldsymbol{\beta}}_1; D_1), \tilde{J}_1 = -\nabla\nabla l(\widetilde{\boldsymbol{\beta}}_1; D_1))$, where $\widetilde{\boldsymbol{\beta}}_1$ is the offline MLE $\hat{\boldsymbol{\beta}}_1$ based on D_1. The dataset D_1 is then assumed to be removed from the hard drive.

At batch $b = 2$, we have access to the entire dataset D_2 and summaries from the previous batch $b = 1$. The loglikelihood at the renewable MLE is $l(\widetilde{\boldsymbol{\beta}}_2; D_2^*) = l(\widetilde{\boldsymbol{\beta}}_2; D_1) + l(\widetilde{\boldsymbol{\beta}}_2; D_2)$.

In the absence of the dataset D_1, the first term requires an approximation:

$$\begin{aligned} l(\widetilde{\boldsymbol{\beta}}_2; D_1) &\approx l(\widetilde{\boldsymbol{\beta}}_1; D_1) + (\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1)^T \nabla l(\widetilde{\boldsymbol{\beta}}_1; D_1) + \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1)^T \nabla\nabla l(\widetilde{\boldsymbol{\beta}}_1; D_1)(\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1) \\ &= l(\widetilde{\boldsymbol{\beta}}_1; D_1) + 0 - \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1)^T \tilde{J}_1 (\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1) \end{aligned}$$

where the last equality is due to $\nabla l(\widetilde{\boldsymbol{\beta}}_1; D_1) = \mathbf{0}$, because the gradient of the loglikelihood evaluated at the MLE is $\mathbf{0}$.

The above suggests an approximation for $l(\widetilde{\boldsymbol{\beta}}_2; D_2^*)$, which we denote by $\tilde{l}_2$:

$$\tilde{l}_2 = l(\widetilde{\boldsymbol{\beta}}_1; D_1) - \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1)^T \tilde{J}_1 (\widetilde{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_1) + l(\widetilde{\boldsymbol{\beta}}_2; D_2),$$

and record $(\widetilde{\boldsymbol{\beta}}_2, \tilde{J}_2, \tilde{l}_2)$.

We next extend the algorithm to a generic bth batch for $b \geq 3$.

For a generic batch $b \geq 3$, we have on record: $(\widetilde{\boldsymbol{\beta}}_{b-1}, \tilde{J}_{b-1}, \tilde{l}_{b-1})$, and the entire bth batch D_b. At batch b, we need to evaluate $l(\widetilde{\boldsymbol{\beta}}_b; D_b^*) = l(\widetilde{\boldsymbol{\beta}}_b; D_{(b-1)}^*) + l(\widetilde{\boldsymbol{\beta}}_b; D_b)$. The second term of the right hand side can be evaluated directly. The first term requires an approximation:

$$\begin{aligned} l(\widetilde{\boldsymbol{\beta}}_b; D_{(b-1)}^*) \approx\ & l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)}^*) + \left(\widetilde{\boldsymbol{\beta}}_b - \widetilde{\boldsymbol{\beta}}_{(b-1)}\right)^T \nabla l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)}^*) \\ & + \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_b - \widetilde{\boldsymbol{\beta}}_{(b-1)})^T \nabla\nabla l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)}^*)(\widetilde{\boldsymbol{\beta}}_b - \widetilde{\boldsymbol{\beta}}_{(b-1)})\,. \end{aligned} \tag{10}$$

Now we approximate each of the three terms on the right hand side of (10).

An approximation of the first term in (10), $l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)}^*)$, is $\tilde{l}_{(b-1)}$, readily available from the record of summaries of past batches. The second term in (10) is close to zero, as

$$\nabla l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)}^*) \approx \nabla l(\hat{\boldsymbol{\beta}}_{(b-1)}; D_{(b-1)}^*) = \mathbf{0}.$$

Concerning the third term in (10),

$$\nabla\nabla l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D^*_{(b-1)}) = \sum_{j=1}^{b-1} \nabla\nabla l(\widetilde{\boldsymbol{\beta}}_{(b-1)}; D_j) \approx \sum_{j=1}^{b-1} \nabla\nabla l(\widetilde{\boldsymbol{\beta}}_j; D_j) = -\tilde{J}_{(b-1)}\,,$$

where $\tilde{J}_{(b-1)}$ is again part of the "historical data" from the previous batch $(b-1)$.

Thus our final approximation for $l(\widetilde{\boldsymbol{\beta}}_b; D^*_b)$ denoted by $\tilde{l}_b$ is

$$\tilde{l}_b = \tilde{l}(\widetilde{\boldsymbol{\beta}}_{b-1}; D^*_{b-1}) - \frac{1}{2}(\widetilde{\boldsymbol{\beta}}_b - \widetilde{\boldsymbol{\beta}}_{b-1})^T \tilde{J}_{b-1}(\widetilde{\boldsymbol{\beta}}_b - \widetilde{\boldsymbol{\beta}}_{b-1}) + l(\widetilde{\boldsymbol{\beta}}_b; D_b).$$

After processing the data in batch b, we record $(\widetilde{\boldsymbol{\beta}}_b, \tilde{J}_b, \tilde{l}_b)$ and the dataset D_b can be from storage.

3.4 BMA With Online Methods for Small p

For any batch b, and a given model $\boldsymbol{\gamma}$, one can approximate the loglikelihood evaluated at the MLE, $l(\hat{\boldsymbol{\beta}}_{b\boldsymbol{\gamma}}; {D_b}^*)$, based on the Online 1 or the Online 2 methods. We denote these estimators by $\check{l}_{b,\boldsymbol{\gamma}}$ and $\tilde{l}_{b,\boldsymbol{\gamma}}$, respectively. Note the notation are extensions from $\check{l}_b$ or $\tilde{l}_b$ to include index of the model under consideration. Equipped with the approximate loglikelihood, one can now approximate the BIC for all models $\boldsymbol{\gamma}$ in the model space $\boldsymbol{\Gamma}$ using (3). Then one can calculate the posterior probabilities of all 2^p models via (2) for BMA. As enumeration of all models is typically not possible for $p > 30$, we provide a method to overcome this difficulty for large p in the next section.

3.5 Addressing the Second Difficulty for Large p via MCMC Sampling for Streaming Data

In this section, we focus on the situation where we would like to explore different models indexed by $\boldsymbol{\gamma}$ and make inference based on online posterior approximations, $\tilde{\pi}(\boldsymbol{\gamma}|D)$, when enumeration of all models in $\mathbf{\Gamma}$ is computationally infeasible. The types of inference may include prediction of the response variable, estimation of the regression coefficients, variable selection and so on. In classical offline settings for model exploration, MCMC can be used to visit the most promising $\boldsymbol{\gamma}$s. The challenge of implementing this when combined with the online set up, is that for our online approximations to work for any model $\boldsymbol{\gamma}$ at time b, the summaries of this particular $\boldsymbol{\gamma}$ from time $t < b$ also need to be available. Since enumeration of all 2^p models in $\mathbf{\Gamma}$ is not possible, a potential solution is we we first select a promising subset of models, and then record the online summary statistics such as MLE, loglikelihood, and Hessian for this chosen subset. This implies inference for BMA will be restricted to this particular pool of models in the subsequent batches. So, it is important to identify a good, sufficiently large candidate pool of models.

In order to provide an online solution to BMA, based on a promising pool of models, we propose a method that consists of two stages:
i) **a screening stage**, in which MCMC sampling is employed to select a good subset of models, and
ii) **a post screening stage**, during which we provide an online approximation for BMA, as in Section 3.4, except that here it is not for the entire model space $\mathbf{\Gamma}$, but the chosen subset of $\mathbf{\Gamma}$, determined in the above screening stage.

Note that in the offline scenario, the MCMC solution runs a Markov chain that has $\tilde{\pi}(\boldsymbol{\gamma}|D)$ as its invariant distribution. A popular MCMC algorithm for Bayesian variable selection is the MCMC model composition, also known as the MC^3 algorithm (Madigan and York, 1995; Raftery et al., 1997; Clyde et al., 2011; Ghosh and Tan, 2015). In the online scenario, let B_0 be a relatively small number of batches, such that the computational cost of the offline MCMC for $D^*_{B_0}$ is not expensive. For each of the time points $b = 1, \ldots, B_0$, we run an offline MC^3 Markov chain, $\{\boldsymbol{\gamma}^{[b](0)}, \boldsymbol{\gamma}^{[b](1)}, \ldots, \boldsymbol{\gamma}^{[b](T)}\}$, with a target distribution $\tilde{\pi}(\boldsymbol{\gamma}|D^*_b)$ analogous to that in (2).

We perform stability checks of the pool of sampled models at batches $b = 1, \cdots, B_0$, respectively, to track their stability across successive batches. When stability is reached the screening stage is ended. Specifically, we choose a metric that measures the stability of predictions. Let $\lambda_i^{[b]}$ denote an MCMC estimate for the mean of the posterior predictive distribution for a future response variable, say $\tilde{Y}_i$. We check the following metric for stability:

$$\sqrt{\frac{1}{N_b}\sum_{i=1}^{N_b}(\lambda_i^{[b]} - \lambda_i^{[b-1]})^2} \tag{11}$$

against a pre-specified threshold. We will discuss the definition of $\lambda_i^{[b]}$ in the context of logistic regression, in more details in the next section on simulation study.

4 Simulation Study for Online BMA for Logistic Regression

We conduct simulation studies that compare the results from the traditional offline BMA method, with the Online 1 and 2 methods, and the full model offline MLE, for $p = 10$ and $p = 50$, which excludes the intercept term. For $p = 10$, we always include the intercept term, and each of the remaining 10 variables can be included or excluded from the model, which leads to a model space with $2^{10} = 1,024$ models. Since this is a reasonably small model space, we can list all $1,024$ models in the computer memory and perform BMA with all models in $\mathbf{\Gamma}$. This is sometimes referred to as an enumerable model space, because all models can be stored in the computer memory. For $p = 50$, the model space $\mathbf{\Gamma}$ has $2^{50} = 1.1259 \times 10^{15}$ models. Since all models cannot be stored, this is called a non-enumerable model space. For both cases, we use a discrete uniform prior, that is $p(\boldsymbol{\gamma}) = 1/2^p$. We describe our simulation study results for these two scenarios, in more detail below.

4.1 Enumerable Model Space

We generate one hundred datasets from the logistic regression model with $p = 10$. The covariates are generated independently from a normal distribution with mean 0 and standard deviation 3. The regression coefficients are set to 0.2 for the first five coefficients including the intercept term and the rest are 0. We consider $B = 100$ batches where the batch size is $n_b = 100$ for each batch. The regression coefficients are estimated using $\widehat{\boldsymbol{\beta}}^{[b]}$, which is the BMA estimate of $\boldsymbol{\beta}$ for batch b. For the bth batch, with aggregated data D_b^*, the BMA estimate is given by

$$\widehat{\boldsymbol{\beta}}^{[b]} = \sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{[b]} \tilde{\pi}(\boldsymbol{\gamma}|D_b^*), (12)$$

where $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{[b]} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{[b]}$ for $\gamma_j = 1$, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{[b]} = 0$ for $\gamma_j = 0$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{[b]}$ is the MLE and $\tilde{\pi}(\boldsymbol{\gamma}|D_b^*)$ is the posterior probability for model $\boldsymbol{\gamma}$, for the bth batch. The offline method uses the offline MLE and posterior probability to compute these quantities. Online 1 and 2 methods use the renewed estimate of MLE (Luo and Song, 2020), and the Online 1 or 2 approximations to the loglikelihood, respectively, proposed by us in the previous section, for approximation of $\tilde{\pi}(\boldsymbol{\gamma}|D_b^*)$. The root mean squared errors (RMSE) for estimating the true regression coefficients are computed for all methods and shown in the top panel of Figure 1. Overall, the online BMA methods have similar RMSE as the gold standard offline BMA. Furthermore, we see an advantage of the second order Taylor expansion used in Online 2, as it tends to have a smaller RMSE than Online 1 throughout. The RMSE of all methods decrease as the batch sizes increase, and the aggregate data size increases, as expected.

For prediction, we generate a new design matrix and new response variables, according to the aforementioned true model with sample size $N_B = 10,000$. We use the following definition of RMSE for prediction:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N_B} (\tilde{Y}_i - \hat{P}(\tilde{Y}_i = 1))^2}{N_B}}, \tag{13}$$

where $\tilde{Y}_i$ is the response variable in the test set of size N_B, and $\hat{P}(\tilde{Y}_i = 1)$ is its BMA estimate from the offline and online methods, for $i = 1, \ldots, N_B$. We also calculate an oracle RMSE which is the best case scenario when an oracle method uses the true $P(\tilde{Y}_i = 1)$ in the data generating model as the estimate. The RMSE for prediction is shown in the bottom panel of Figure 1. We observe a similar trend as the plot in the top panel: both online methods are good approximations of the offline method, with Online 2 being a better approximation consistently. All methods approach the oracle RMSE value as the batch size increases. Running times for offline and online Bayesian methods for a single replicate, and looping through all batches $1, 2, \ldots, B$, are approximately 1273 and 28 seconds respectively, that is **the running time of the offline method is about 45 times that of the online methods**. The full model offline MLE is less accurate than the BMA methods.

4.2 Non-enumerable Model Space

We now generate one hundred datasets from the logistic regression model with $p = 50$. The covariates are generated independently from a normal distribution with mean 0 and standard deviation $\sqrt{3}$. The intercept is 0.1, the next 15 regression coefficients are 0.2, and the rest are 0. We consider $B = 50$ batches where the batch size $n_b = 200$ for each batch. Since for $p > 30$, all 2^p models cannot be stored in the computer memory, we conduct a screening to choose a promising pool of models. Suppose the binary response variable in logistic regression takes the values 1 and 0, denoting success and failure, respectively. The metric $\lambda_i^{[b]}$ was introduced in (11) for checking stability in the screening stage. We now discuss methods to define $\lambda_i^{[b]}$ in the context of logistic regression. Here $\lambda_i^{[b]}$ is an estimate of $\mathrm{E}(\tilde{Y}_i | D_b^*) = P(\tilde{Y}_i = 1 | D_b^*)$, where $\tilde{Y}_i$ is a future response variable, generated from the same process that generated Y_i. We use a plug-in estimate for this probability: $\exp(\boldsymbol{x}_i \overline{\boldsymbol{\beta}}^{[b]})/(1 + \exp(\boldsymbol{x}_i \overline{\boldsymbol{\beta}}^{[b]}))$, where $\overline{\boldsymbol{\beta}}^{[b]} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\beta}^{[b](t)}$ is the MCMC estimate of $\boldsymbol{\beta}$, and $\boldsymbol{\beta}^{[b](t)}$ is the $(p+1)$-dimensional esti-

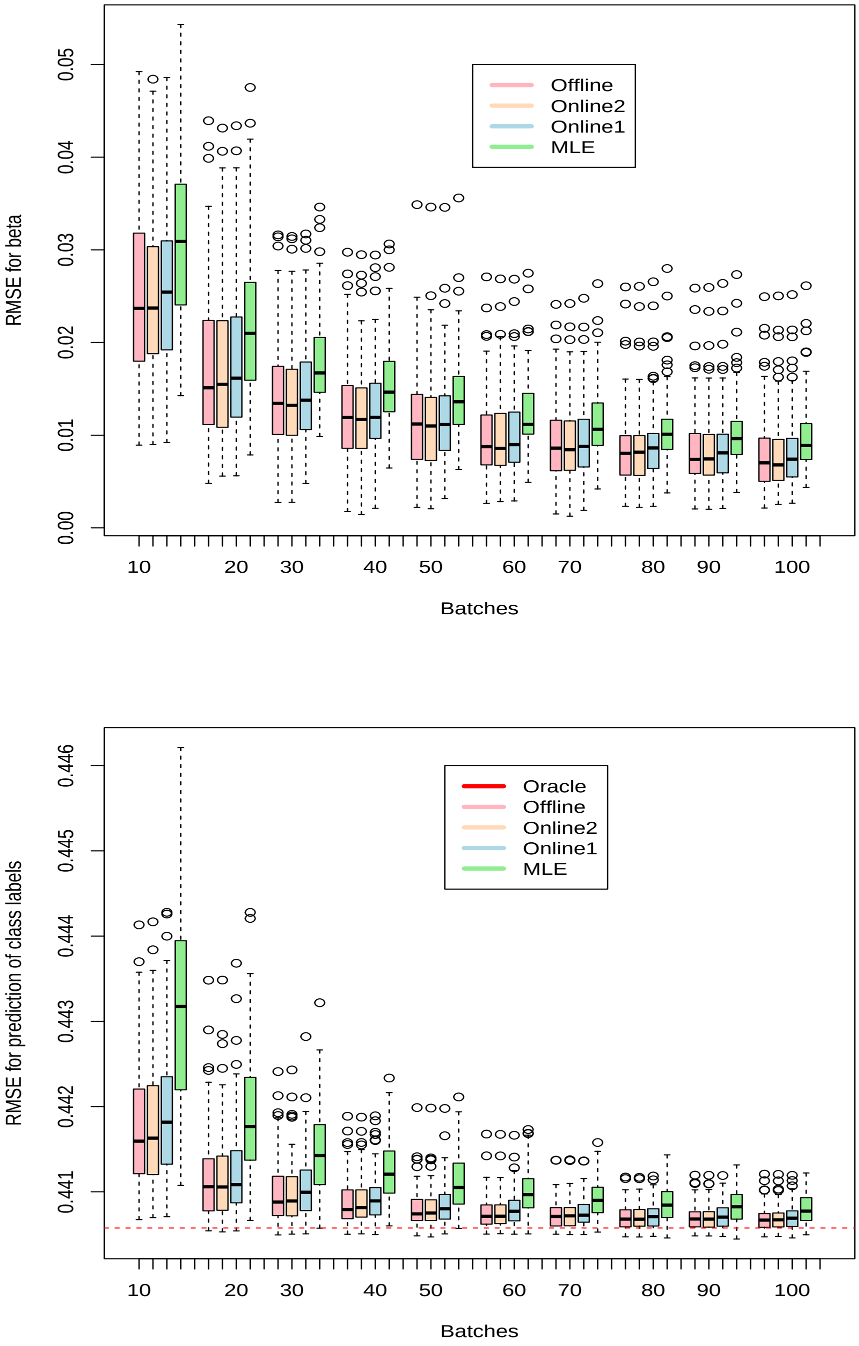

Figure 1: Performance of offline, Online 2, Online 1 BMA methods, and the full model offine MLE for $p = 10$, for batches $10, 20, \ldots, 100$. **Running time of the offline Bayesian method is about 45 times that of the online Bayesian methods**.

mate at iteration t. This plug-in estimate is computationally inexpensive to calculate. But if one is willing to spend more computing effort, other Bayesian estimates could be used. In the screening stage we run the offline MC^3 algorithm for 10,000 iterations and monitor the stability metric defined in (11). Screening is stopped if one of the two conditions is met: the metric reaches a threshold of 0.02 or less, or the 10th batch is reached. Offline computations become prohibitively slow as the number of batches increases, so screening is stopped after batch 10. Essentially these are tuning parameters for screening, which need to be set by the data analyst, depending on the size of the datasets and speed of the algorithms. Online methods are implemented with the distinct pool of models sampled in the final screening stage. Figure 2 shows the results, which are similar to the enumerable case. Here the offline method needs 128.4 minutes for a single dataset, looping through all batches, and the online methods take 4.6 minutes for the same. **The running time of the offline method is about 28 times that of the online methods.** The drop in accuracy of the MLE is quite pronounced here.

5 Application

We apply the methods developed in the previous sections to streaming data from the National Automotive Sampling System (NASS) Crashworthiness Data System (CDS). We analyze occupant level data on traffic crashes from 2009 to 2015 (Luo and Song, 2020). The response variable is an indicator for injury, which takes the value 1 if the occupant had an injury and 0 otherwise. This is based on the variable MAIS (maximum known abbreviated injury scale) for this occupant. We consider the following predictors related to the occupant: AGE, HEIGHT, OCCRACE (occupant's race), OCETHNIC (occupant's ethnicity), ROLE

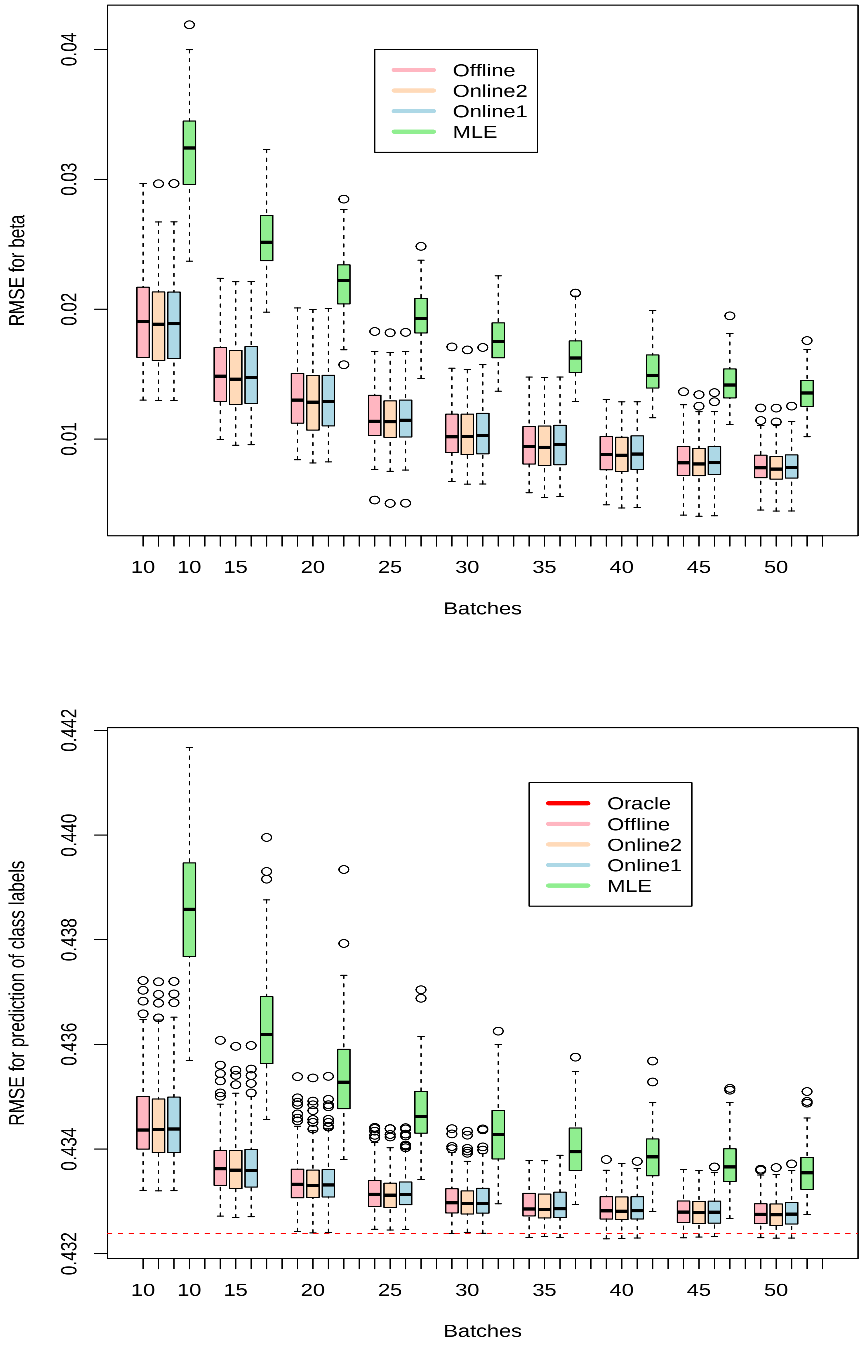

Figure 2: Performance of offline, Online 2, Online 1 BMA methods, and the full model offline MLE for $p = 50$, for batches $10, 15, \ldots, 50$. **Running time of the offline Bayesian method is about 28 times that of the online Bayesian methods**.

(driver vs. passenger), SEX, WEIGHT, and other predictors related to the vehicle and the accident: BAGFAIL (air bag system failure), BELTANCH (shoulder belt upper anchorage adjustment), EJCTAREA (ejection area), and PARUSE (police reported restraint use). Here AGE, HEIGHT, and WEIGHT are numeric predictors and the rest are categorical, leading to a total of 33 predictors, including the intercept term. To demonstrate the utility of BMA in the presence of noise, we add 300 noise variables generated from a normal distribution with mean 0 and variance 9, with a pairwise correlation of 0.5, following the idea of Ghosh and Reiter (2013).

The dataset exhibits separation, that is for some linear combination of the predictors, a subset of the response variables can be predicted perfectly. The concept of separation was introduced by Albert and Anderson (1984), who showed that the MLE does not exist in such a scenario. The effect of the choice of the prior distribution on the posterior distribution, in binary regression models, under separation, has been studied extensively (Chen and Shao, 2001; Gelman et al., 2008; Speckman et al., 2009; Ghosh et al., 2018; Ghosh, 2019). Proper priors are necessary in this case to guarantee a proper posterior. While there are many choices of proper priors, we use the prior recommended by Greenland and Mansournia (2015), as this prior has an interpretation of augmenting rows to the design matrix and the vector of response variables, which in turn makes computation of the posterior mode straightforward, with standard software for calculating the MLE.

Greenland and Mansournia (2015) suggest the use of the $\log F(m, m)$ prior for the regression coefficients in logistic regression under separation. This prior has lighter tails than a t-prior with m degrees of freedom but heavier tails than a normal. Suppose we use this prior for β_j given $\gamma_j = 1$, then the prior density is given as follows (Brown et al., 2002):

$$
\begin{aligned}
p(\beta_j|\gamma_j = 1) &= \frac{1}{B(m/2, m/2)} \frac{e^{\beta_j m/2}}{(1+e^{\beta_j})^m} \\
&\propto \left(\frac{e^{(0\beta_0 + \ldots 0\beta_{j-1} + 1\beta_j + 0\beta_{j+1} + \ldots 0\beta_p)}}{1 + e^{(0\beta_0 + \ldots 0\beta_{j-1} + 1\beta_j + 0\beta_{j+1} + \ldots 0\beta_p)}} \right)^{m/2} \times \\
&\quad \left(\frac{1}{1 + e^{(0\beta_0 + \ldots 0\beta_{j-1} + 1\beta_j + 0\beta_{j+1} + \ldots 0\beta_p)}} \right)^{m/2}, \qquad (14)
\end{aligned}
$$

where $B(.,.)$ is a Beta function. The last two lines in equation (14) show that the prior can be written in the same form as the likelihood, for observations with $\boldsymbol{x}_i^T = (0, \ldots, 0, 1, 0, \ldots, 0)$, with 1 in the $(j+1)$th position, $j = 0, 1, \ldots, p$. The prior represents m observations with $m/2$ successes and $m/2$ failures. Thus for computation of the posterior mode, the m rows corresponding to the prior can be appended to the data and standard software for maximizing the loglikelihood will then effectively maximize the log-posterior, up to a normalizing constant. In principle one could consider independent priors of this form on all the components of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, but for a large p, that can potentially add excessive prior information. Since we mainly use this prior to address separation, we first identify which variables are involved in separation via the R package `detectseparation` (Kosmidis et al., 2023), and then specify this prior only for the predictors affected by separation. We set $m = 4$ following the suggestion of Agresti and Coull (1998), for binomial proportions. Under separation, Ghosh et al. (2018) have also recommended using priors that have less heavy tails than a Cauchy, as it can lead to unusually large regression coefficients in logistic regression. For $m = 4$, the $\log F(m, m)$ prior will have less heavy tails than a t with 4 degrees of freedom, and thus much lighter than a Cauchy prior, which seems reasonable in this scenario. We find that 11 of the $(p+1) = 333$ predictors, are involved in separation, which leads to augmenting 44 rows to the first batch, corresponding to the prior.

Here we consider $B = 20$ batches, with batch size $n_b = 800$, for each batch, leading to $N_B = 16,000$. We keep the last 438 observations for out of sample prediction. Let ρ denote the prior inclusion probability, that is $\rho = P(\gamma_j = 1)$, for $j = 1, \ldots, p$. In the simulation studies, we set $\rho = 1/2$. Since $p = 332$ is relatively large compared to the number of predictors in the simulation studies, here we put a prior on ρ. Specifically, we let $\rho \sim \text{Beta}(a, b)$, with a commonly adopted choice of $a = b = 1$, which leads to a continuous uniform prior on ρ (Scott and Berger, 2010). We compare the offline and online Bayesian methods, as well as the posterior mode based on the full model. Screening is performed as in Section 4.2, to choose the set of models. Here the screening period ends after the 5th batch, and online computations start from the 6th batch onwards. The MCMC is run for 100,000 iterations for all methods (offline and online in screening stage).

Figure 3 shows the results for out of sample prediction of class labels, evaluated using equation (13). Here the offline method needs 14.54 hours and the online methods take 0.8 hour for 20 batches. **Thus, the running time of the offline method is about 18 times that of the online methods.** The advantage of using methods based on BMA over the full model is most pronounced in the early batches, and a distinct gain from BMA is evident throughout.

The estimated posterior inclusion probabilities for offline vs. online methods after analyzing the 20th batch, are displayed in Figure 4. The median probability models (MPMs) (Barbieri and Berger, 2004) based on the offline method and Online 2 are in close agreement. The MPM based on the offline method includes AGE, one category for SEX, WEIGHT, all categories for PARUSE, one category of EJCTAREA. Online 2 includes all the variables as offline, except one category of PARUSE, which indicates the use of shoulder belt. Online 1 includes all the variables as Online 2, and additionally includes one category of OCETHNIC, which has an inclusion probability closer to 0.5 across all methods. All the noise variables are effectively dropped by all methods.

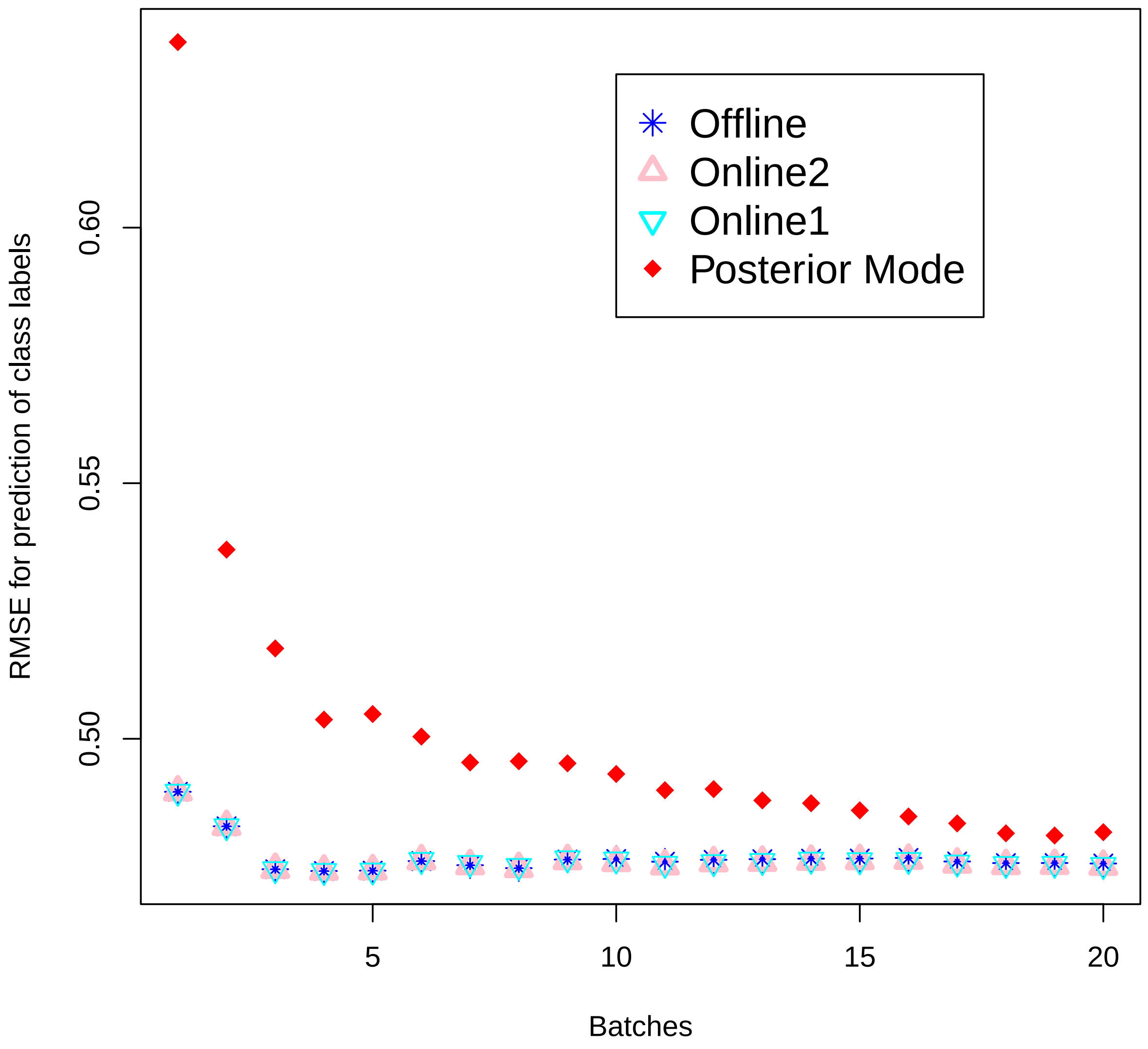

Figure 3: Performance of offline, Online 2, and Online 1 BMA methods, and the full model offline posterior mode for the application with $p = 332$, for batches $1, 2, \ldots, 20$. **Running time of the offline Bayesian method is about 18 times that of the online Bayesian methods**.

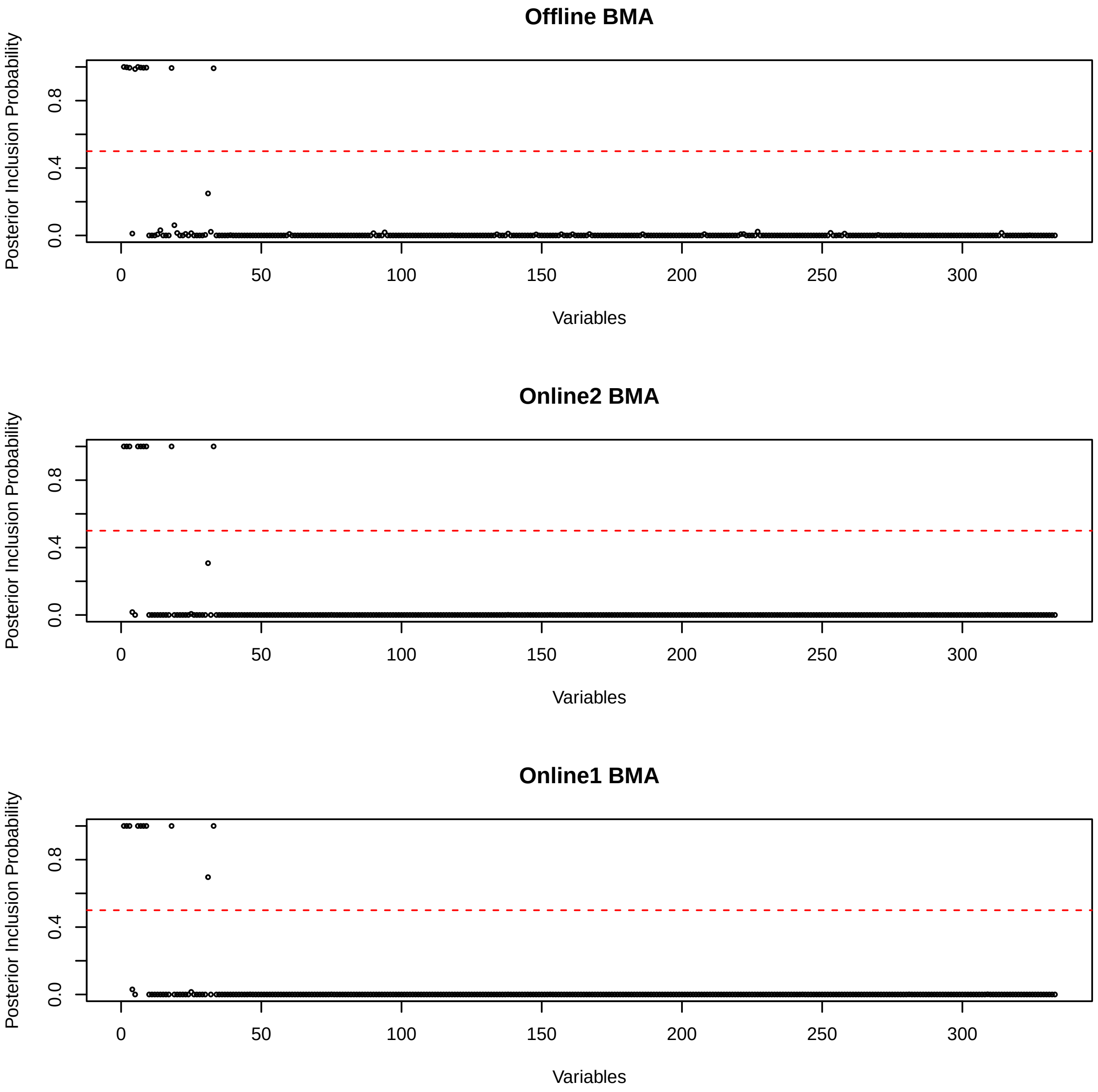

Figure 4: Estimates of posterior inclusion probabilities from offline, Online 2, and Online 1 BMA methods, for the application with $p = 332$, for the 20th batch. The dashed red line is at 0.5, the prior inclusion probability.

6 Discussion

In the context of streaming data, when vast amounts of data can arrive sequentially, data storage becomes expensive and traditional or offline methods become progressively slower with accumulation of more and more data. Online BMA methods introduced in this article aim to address these two challenging aspects of streaming data, by storing only summaries of historical data, and using a technique called renewable estimation, where the most recent estimates are renewed with the current batch of data. To the best of our knowledge, the online BMA methods proposed in this article, are one of the first to exploit the renewable estimation framework, to incorporate variable selection uncertainty in the context of GLMs. The results based on simulation studies and a real data analysis on traffic crashes suggest that our online BMA methods can offer a fast and effective way to address model uncertainty, with accuracy comparable to the gold standard offline BMA methods.

There are many possible variations and extensions of the methods developed in this paper. Our current implementation of screening for large p involves running offline MCMC for several batches in the initial period, to choose the pool of important models. While this method is appealing and works quite well in practice, it comes at the cost of some computational burden. One possibility is to first reduce the number of predictors by methods such as sure screening (Fan and Lv, 2008), and then employ our current method of screening using MCMC, which can be much faster. Alternative choices are available for the metric used to assess stability in the screening stage, such as those based on the estimated regression coefficients or marginal posterior inclusion probabilities, instead of classification probabilities.

Our focus has been on point estimates for BMA in logistic regression, but if one is interested in credible intervals for classification probabilities, those can be obtained via Monte Carlo sampling from the approximate joint posterior distribution of models and regression parameters. For illustration, we used logistic regression, but the methods are applicable to other members of the GLM family, such as Poisson and gamma regression.

Our implementation of the BIC approximation to the marginal likelihood is based on the MLE of $\boldsymbol{\beta}_\gamma$ or the posterior mode, under data augmentation priors (Greenland and Mansournia, 2015). Alternative priors could be placed such as normal, power priors (Ibrahim et al., 2015), using the method of Raftery (1996).

Our current approach for large p relies on screening using MCMC, to choose a pool of models with high posterior probability. The online methods are implemented for this pool for subsequent batches. Developing an online sampling algorithm for models, that is letting the pool of models vary with batches, based on MCMC, variational Bayesian methods, or other stochastic algorithms such as the shotgun stochastic search (Hans et al., 2007) is an interesting direction of research. Other interesting avenues include extension to models that are not covered under the current framework.

Acknowledgments

The authors thank Drs. Kate Cowles and Amy Herring for constructive comments on a previous version of this manuscript.

References

Agresti, A. (2013). *Categorical Data Analysis*. Wiley. 6

Agresti, A. and Coull, B. A. (1998). “Approximate Is Better than ”Exact” for Interval Estimation of Binomial Proportions.” *The American Statistician*, 52(2): 119–126. 21

Albert, A. and Anderson, J. A. (1984). “On the Existence of Maximum Likelihood Estimates in Logistic Regression Models.” *Biometrika*, 71(1): 1–10. 20

Barbieri, M. and Berger, J. (2004). “Optimal predictive model selection.” *Annals of Statistics*, 32(3): 870–897. 22

Brown, B. W., Spears, F. M., and Levy, l. B. (2002). “The log F: A Distribution for All Seasons.” *Computational Statistics*, 17: 47–58. 21

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480.
URL http://biomet.oxfordjournals.org/cgi/content/abstract/asq017v1 6

Chen, M.-H. and Shao, Q.-M. (2001). "Propriety of Posterior Distribution for Dichotomous Quantal Response Models." *Proceedings of the American Mathematical Society*, 129(1): 293–302. 20

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). "Bayesian adaptive sampling for variable selection and model averaging." *Journal of Computational and Graphical Statistics*, 20(1): 80–101. 13

Fan, J. and Lv, J. (2008). "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911. 25

Gelman, A., Jakulin, A., Pittau, M., and Su, Y. (2008). "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics*, 2(4): 1360–1383. 20

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88: 881–889. 6

— (1997). "Approaches for Bayesian variable selection." *Statistica Sinica*, 7: 339–374. 6

Ghosh, J. (2019). "Cauchy and other shrinkage priors for logistic regression in the presence of separation." *WIREs Computational Statistics*, 11(6): e1478. 20

Ghosh, J., Li, Y., and Mitra, R. (2018). "On the use of Cauchy prior distributions for Bayesian logistic regression." *Bayesian Analysis*, 13: 359–383. 20, 21

Ghosh, J. and Reiter, J. P. (2013). "Secure Bayesian model averaging for horizontally partitioned data." *Statistics and Computing*, 23(3): 311–322. 20

Ghosh, J. and Tan, A. (2015). "Sandwich algorithms for Bayesian variable selection." *Computational Statistics and Data Analysis*, 81: 76–88. 13

Greenland, S. and Mansournia, M. A. (2015). "Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions." *Statistics in Medicine*, 34(23): 3133–3143. 20, 25

Han, R., Luo, L., Lin, Y., and Huang, J. (2024). "Online inference with debiased stochastic gradient descent." *Biometrika*, 111(1): 93–108. 3

Hans, C., Dobra, A., and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, 102: 507–516. 6, 26

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial (with discussion)." *Statistical Science*, 14(4): 382–401. 4

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). "The power prior: theory and applications." *Statistics in Medicine*, 34: 3724–3749. 25

Jørgensen, B. (1987). "Exponential dispersion models." *Journal of the Royal Statistical Society, Series B*, 49: 127–162. 6

Kosmidis, I., Schumacher, D., and Schwendinger, F. (2023). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.3. URL https://github.com/ikosmidis/detectseparation 21

Liu, F., Chakraborty, S., Li, F., Liu, Y., and Lozano, A. C. (2014). "Bayesian regularization via graph laplacian." *Bayesian Analysis*, 9(2): 449 – 474.
URL https://doi.org/10.1214/14-BA860 6

Luo, L., Han, R., Lin, Y., and Huang, J. (2023). "Online inference in high-dimensional generalized linear models with streaming data." *Electronic Journal of Statistics*, 17(2): 3443–3471. 3

Luo, L. and Song, P. X. (2020). "Renewable estimation and incremental inference in generalized linear models with streaming data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82. 3, 7, 8, 15, 18

Madigan, D. and York, J. (1995). "Bayesian graphical models for discrete data." *International Statistical Review*, 63: 215–232. 6, 13

McCormick, T. H., Raftery, A. E., Madigan, D., and Burd, R. S. (2012). "Dynamic logistic regression and dynamic model averaging for binary classification." *Biometrics*, 68(1): 23–30. 3

Nie, L. and Ročková, V. (2023). "Bayesian Bootstrap Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 118(543): 2013–2028. 6

Onorante, L. and Raftery, A. E. (2016). "Dynamic model averaging in large model spaces using dynamic Occams window." *European Economic Review*, 81: 2–14. Model Uncertainty in Economics.
URL https://www.sciencedirect.com/science/article/pii/S0014292115001099 3

Raftery, A. E. (1996). "Approximate Bayes factors and accounting for model uncertainty in generalised linear models." *Biometrika*, 83: 251–266. 5, 26

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92: 179–191. 6, 13

Robbins, H. and Monro, S. (1951). "A stochastic approximation method." *The Annals of Mathematical Statistics*, 400–407. 3

Sato, M.-A. (2001). "Online model selection based on the variational Bayes." *Neural computation*, 13(7): 1649–1681. 3

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). "Online updating of statistical inference in the big data setting." *Technometrics*, 58(3): 393–403. PMID: 28018007.
URL https://doi.org/10.1080/00401706.2016.1142900 3

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multicplicity adjustment in the variable-selection problem." *Annals of Statistics*, 38(5): 2587–32619. 22

Shi, C., Song, R., Lu, W., and Li, R. (2021). "Statistical inference for high-dimensional models via recursive online-score estimation." *Journal of the American Statistical Association*, 116(535): 1307–1318. 3

Speckman, P. L., Lee, J., and Sun, D. (2009). "Existence of the MLE and Propriety of Posteriors for a General Multinomial Choice Model." *Statistica Sinica*, 19: 731–748. 20

Toulis, P., Airoldi, E., and Rennie, J. (2014). "Statistical analysis of stochastic gradient methods for generalized linear models." In *International Conference on Machine Learning*, 667–675. PMLR. 3

Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2016). "Statistical methods and computing for big data." *Statistics and Its Interface*, 9: 399–414. 3

Williams, J., Xu, S., and Ferreira, M. (2023). "BGWAS: Bayesian variable selection in linear mixed models with nonlocal priors for genome-wide association studies." *BMC Bioinformatics*, 24. 6

Zhang, T. and Yang, B. (2021). "Online multiple learning with working sufficient statistics for generalized linear models in big data." *Statistics and Its Interface*, 14(4): 403–416. 3