Subsampled Information Criteria for Bayesian Model Selection in the Big Data Setting

Lijiang Geng* Yishu Xue* Guanyu Hu*

Abstract

Bayesian methods face unprecedented challenges in the era of big data, as the evaluation of likelihood in each iteration is computationally intensive. To deal with this bottleneck, recent literatures focus mostly on speeding up Markov chain Monte Carlo (MCMC). Model selection, which is an important topic, has not received much attention. In the Bayesian context, deviance-based criteria, such as the deviance information criterion (DIC), are well-known for model selection purposes. In this article, we introduce the subsampled DIC and the subsampled information criterion IC_{AT} in the big data context. Under reasonable regularity conditions, we show that our proposed subsampled criteria closely approximate their full data counterparts. Extensive simulation studies are conducted to evaluate the empirical performance of the proposed criterion. The usage of our proposed criterion is further illustrated with the analysis of two datasets, a household income data from the 1994 Census, and the Covertype Data Set.

Key Words: DIC, IC_{AT}, MCMC, Nonuniform Subsample

1. Introduction

Since the 1990s, Bayesian methods have become increasingly popular due to the introduction of powerful sampling algorithms such as the Markov Chain Monte Carlo (MCMC). With the advancement in and prevalence of computer technology, however, big data poses challenges to Bayesian methods. Evaluation of the full data likelihood function in each iteration makes the MCMC time consuming when the number of observations is large. Towards this end, two major approaches aiming at speeding up MCMC have been proposed, the parallel MCMC (Wang and Dunson, 2013), and the subsampled MCMC (Bardenet et al., 2014, 2017; Quiroz et al., 2018; Hu and Wang, 2018). Most of these literatures, however, focus on Bayesian estimation, while less attention has been paid to Bayesian model selection.

In the Bayesian framework, the deviance information criterion (DIC; Spiegelhalter et al., 2002) is one of the most frequently used tools for model selection. In addition, there are many other deviance based criteria, e.g., the Bayesian Predictive Information Criterion (BPIC; Ando, 2007), the Bayesian predictive distribution-based information Criterion (IC_{AT}; Ando and Tsay, 2010) and the Posterior Averaging Information Criterion (PAIC; Zhou, 2011). For most deviance based information criteria, calculation of the likelihood deviance and posterior deviance based on full data is required in every MCMC iteration. In this work, we introduce the subsample DIC and subsampled IC_{AT} based on the Bayesian predictive distribution, which are calculated using a small portion of full data in every MCMC iterations for Bayesian model selection, yet still produces credible model selection results. Similar to in Hu and Wang (2018), we use nonuniform probabilities in drawing the subsamples, which requires even smaller subsample sizes to achieve the same approximation accuracy than using uniform sampling.

The rest of this paper is organized as follows. In Section 2, we review the subsampled MCMC algorithm, the DIC and the IC_{AT}. In Section 3, we propose the subsampled DIC and the subsampled IC_{AT}, and present their theoretical properties. Simulation study

*Department of Statistics, University of Connecticut, Storrs, CT 06269

Algorithm 1 Metropolis–Hastings algorithm

```
for $k \leftarrow 1$ to $N$ do
    $\theta \leftarrow \theta_{k-1}$
    $\theta' \sim q(. \mid \theta)$
    $u \sim U(0, 1)$
    $\alpha = \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}$
    if $\alpha > u$ then
        $\theta_k \leftarrow \theta'$ {Accept}
    else
        $\theta_k \leftarrow \theta$ {Reject}
    end if
    Return $\theta_k$, $k = 1, ..., N$
end for
```

results are shown in Section 4. In Section 5, two applications of our proposed methods are performed, one using the Covertype Data (Collobert et al., 2002), and the other using household income data extracted from the 1994 Census database. We conclude with a brief discussion in Section 6. For ease of exposition, additional technical results are given in the appendix.

2. Background and Related Work

Consider a dataset $\boldsymbol{x} = \{x_1, ..., x_n\}$ with associated likelihood $p(\boldsymbol{x} \mid \theta)$ where $\theta \in \Theta$ is the underlying distribution parameter. We assume that the observations are conditionally independent given the value of θ, i.e., $p(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} p(x_i|\theta)$. Given a prior distribution $\pi_0(\theta)$, Bayesian inference often relies on the posterior distribution of θ, that is,

$$\pi(\theta) = \frac{p(\boldsymbol{x} \mid \theta) \times \pi_0(\theta)}{\int p(\boldsymbol{x} \mid \theta) \times \pi_0(\theta) \mathrm{d}\theta} \propto \pi_0(\theta) \prod_{i=1}^{n} p(x_i \mid \theta), \tag{1}$$

where $\pi(\theta)$ denotes the posterior distribution of θ. As $\pi(\theta)$ is intractable in most cases, MCMC methods are often used to generate samples from the posterior distribution for statistical inference.

2.1 Subsampled MCMC

The Metropolis–Hastings (MH) algorithm is a widely used method in Bayesian analysis to sample approximately form $\pi(\theta)$. It proposes a candidate parameter value θ' from the proposal distribution $q(.|\theta)$, and accept θ' probabilistically based on the acceptance probability α. For ease of discussion, we present the standard MH algorithm in Algorithm 1. The standard MH algorithm evaluates $\pi(.)$ for θ and θ' in each iteration, which means the full data likelihood needs to be calculated per iteration. When datasets are large, this could be very computationally intensive. To scale up the MH algorithm, Korattikara et al. (2014) proposed to make the acceptance decision based on a sequential hypothesis test using uniform subsamples adaptively, and control the false acceptance probability; Bardenet et al. (2014) proposed to adaptively approximate the full data likelihood using uniform subsamples under controlled acceptance decision error, and Quiroz et al. (2018) proposed a two-stage delayed acceptance approach. Approximated full data likelihood based on uniform subsamples is used in the first stage to get a approximate acceptance decision, and if the

Algorithm 2 Nonuniform subsampled Metropolis-Hastings algorithm

```
for $k \leftarrow 1$ to $N$ do
    $\theta \leftarrow \theta_{k-1}$
    $\theta' \sim q(.|\theta)$
    $u \sim U(0, 1)$
    $\psi(u, \theta, \theta') \leftarrow \frac{1}{n} \log \left( u \frac{\pi_0(\theta) q(\theta'|\theta)}{\pi_0(\theta') q(\theta|\theta')} \right)$
    $x_1^*, \cdots, x_r^* \overset{\eta}{\sim} X$ {Subsample with replacement according to $\eta_1, ..., \eta_n$}
    $\ell_r^*(\theta) \leftarrow \frac{1}{r} \sum_{i=1}^{r} \frac{\log\{p(x_i^*|\theta)\}}{n\eta_i^*}$
    $\ell_r^*(\theta') \leftarrow \frac{1}{r} \sum_{i=1}^{r} \frac{\log\{p(x_i^*|\theta')\}}{n\eta_i^*}$
    $\Lambda^*(\theta, \theta') \leftarrow \ell_r^*(\theta') - \ell_r^*(\theta)$
    if $\Lambda^*(\theta, \theta') > \psi(u, \theta, \theta')$ then
        $\theta_k \leftarrow \theta'$ {Accept}
    else
        $\theta_k \leftarrow \theta$ {Reject}
    end if
    Return $\theta_k$, $k = 1, ..., N$
end for
```

candidate draws are accepted, the final acceptance decision is then made based on full data in the second stage.

The subsampling scheme in Bardenet et al. (2014) and Korattikara et al. (2014) is uniform subsampling, where each data point has equal probability $1/n$ to be sampled without replacement, while Quiroz et al. (2018) uses uniform subsampling with replacement in order to get independent subsamples and better theoretical performance. These two schemes, however, are approximately equivalent when the subsample size $r \ll n$. To further improve the efficiency of the MH algorithm, Hu and Wang (2018) proposed to draw subsamples using nonuniform probabilities $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_n\}$ such that $0 < \eta_i < 1,\ i = 1, ..., n$ and $\sum_{i=1}^{n} \eta_i = 1$, so an approximation of $\ell_n(\theta)$, the full data log likelihood, is

$$\ell_r^*(\theta) = \frac{1}{r} \sum_{i=1}^{r} \frac{1}{n\eta_i^*} \log\{p(x_i^* \mid \theta)\}. \tag{2}$$

It is direct that $\ell_r^*(\theta)$ is an unbiased estimator for $\ell_n(\theta)$. To better approximate $\ell_n(\theta)$, Hu and Wang (2018) adopted nearly optimal rule and proposed an explicit solution for $\boldsymbol{\eta}$ as

$$\eta_i = \frac{|\log\{p(x_i \mid \widehat{\theta})\}|}{\sum_{j=1}^{n} |\log\{p(x_j \mid \widehat{\theta})\}|}, \quad i = 1, \ldots, n, \tag{3}$$

where $\widehat{\theta} = \arg\max_\theta \sum_{i=1}^{n} \log\{p(x_i \mid \theta)\}/n$ is the maximum likelihood estimate of θ. This gives rise to the nonuniform subsampled MH algorithm in Algorithm 2.

2.2 The Deviance Information Criterion

In this section, we briefly review the deviance information criterion (DIC; Spiegelhalter et al., 2002). Based on the DIC, a candidate model with smaller DIC value is favored. The DIC is defined based on the following deviance function:

$$\text{Dev}(\theta) = -2 \log L(\boldsymbol{x} \mid \theta), \tag{4}$$

where $\boldsymbol{x}$ is the observed data and θ is parameter of the likelihood. Based on the deviance in (4), we have the following form of DIC:

$$\text{DIC} = \overline{\text{Dev}(\theta)} + p_D, \tag{5}$$

where $\overline{\text{Dev}(\theta)}$ is a Bayesian measure of model fit defined as the posterior expectation of the deviance, i.e., $\overline{\text{Dev}(\theta)} = \text{E}_{\theta|\boldsymbol{x}}[-2\log L(\boldsymbol{x} \mid \theta)]$. The second term p_D measures the model complexity, which is defined as the difference between the posterior mean of the deviance and the deviance of the posterior mean of the parameter:

$$p_D = \overline{\text{Dev}(\theta)} - \text{Dev}(\widehat{\theta}),$$

where $\widehat{\theta}$ is the posterior mean. The DIC can be written in two other equivalent forms:

$$\text{DIC} = \text{Dev}(\widehat{\theta}) + 2p_D,$$

and

$$\text{DIC} = 2\overline{\text{Dev}(\theta)} - \text{Dev}(\widehat{\theta}).$$

2.3 The Information Criterion (IC$_{\text{AT}}$) Based on Bayesian Predictive Distribution

From a purely Bayesian viewpoint, the Bayesian predictive distribution is a predictive distribution which is invariant to reparameterization. Following the definition in Ando and Tsay (2010) and Spiegelhalter et al. (2014), the Bayesian predictive distribution is

$$p(\boldsymbol{x}_{\text{rep}} \mid \boldsymbol{x}) = \int p(\boldsymbol{x}_{\text{rep}} \mid \theta)p(\theta \mid \boldsymbol{x})\text{d}\theta,$$

where $\boldsymbol{x}_{\text{rep}}$ is replicate data independently generated by the same mechanism of observed data $\boldsymbol{x}$. The Kullback–Leibler (KL) divergence based on the Bayesian predictive distribution is:

$$\text{KL}[g(\boldsymbol{x}_{\text{rep}}), p(\boldsymbol{x}_{\text{rep}} \mid \boldsymbol{x})] = \text{E}_{\boldsymbol{x}_{\text{rep}}}[\log g(\boldsymbol{x}_{\text{rep}})] - \text{E}_{\boldsymbol{x}_{\text{rep}}}[\log p(\boldsymbol{x}_{\text{rep}} \mid \boldsymbol{x})], \tag{6}$$

where $g(\boldsymbol{x})$ is the data generation process of $\boldsymbol{x}$. Based on Equation (6), Ando and Tsay (2010) defined the following information criterion:

$$\text{IC}_{\text{AT}} = -2\log p(\boldsymbol{x} \mid \boldsymbol{x}) + \text{tr}\left[J^{-1}(\widehat{\theta})I(\widehat{\theta})\right], \tag{7}$$

where

$$\widehat{\theta} = \arg\max_{\theta}\{2\log p(\boldsymbol{x} \mid \theta) + \log \pi_0(\theta)\}, \tag{8}$$

and the matrices I and J are given by

$$I(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial\phi(x_i,\theta)}{\partial\theta}\frac{\partial\phi(x_i,\theta)}{\partial\theta^\top}\right), \tag{9}$$

and

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2\phi(x_i,\theta)}{\partial\theta\partial\theta^\top}\right), \tag{10}$$

where $\phi(x_i, \theta) = \log(p(x_i \mid \theta)) + \log\{\pi_0(\theta)/(2n)\}$, and $\pi_0(\theta)$ is the prior distribution of θ. Based on the MCMC output, $-2\log p(\boldsymbol{x} \mid \boldsymbol{x})$ is estimated by:

$$-2\log\left\{\frac{1}{B}\sum_{b=1}^{B} p(\boldsymbol{x} \mid \theta^b)\right\},$$

where $\theta^1, \cdots, \theta^B$ are B effective random samples drawn from the posterior distribution.

3. Subsampled Bayesian Model Assessment Criterion

In this section, we will discuss two subsampled Bayesian model assessment criteria, DIC_{sp} and IC_{sp} and their theoretical properties.

3.1 The Subsampled DIC

The goal of subsampled DIC is to get an approximation of DIC based on subsampled data. From the definition of DIC, its two components, $\overline{\text{Dev}(\theta)}$ and $\text{Dev}(\widehat{\theta})$, need to be approximated.

Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be n independent observations from the cumulative distribution function $F(\widetilde{x})$ with probability density function $p(\widetilde{x})$, $\eta_1, \cdots, \eta_n$ be sampling probabilities, respectively, and θ be the unknown parameter for $p(\widetilde{x})$. The subsampled DIC is defined as:

$$\text{DIC}_{\text{sp}} = -4\text{E}_{\theta|\boldsymbol{x}}[\ell_r(\theta)] + 2\sum_{i=1}^{n} \ell(x_i \mid \widehat{\theta}), \tag{11}$$

where $\widehat{\theta}$ is the posterior mean based on MCMC samples, $\ell_r(\theta)$ is the subsampled log-likelihood approximation. For the b-th MCMC iteration, $\ell_r(\theta^b) = \sum_{i=1}^{r} \frac{\log\{p(x_i^*|\theta^b)\}}{n\eta_i^*}/r$, where $x_1^*, \cdots, x_r^*$ are r subsampled observations with respective subsample probabilities $\eta_1^*, \cdots, \eta_r^*$, the expectation $E_{\theta|\boldsymbol{x}}(\ell_r(\theta))$ is estimated by:

$$E_{\theta|\boldsymbol{x}}\left[\ell_r(\theta)\right] = \frac{1}{B}\sum_{b=1}^{B} \ell_r(\theta^b). \tag{12}$$

Remark 1. For the uniform subsampled methods, $\pi_i = 1/n$, $i = 1, \cdots, n$, the subsampled log likelihood is $\ell_r(\theta^b) = \sum_{i=1}^{r} \log\{p(x_i^* \mid \theta^b)\}/r$.

3.2 The Subsampled IC_{AT}

For notation simplicity, we denote IC_{AT} in (7) as IC in this section. In order to approximate the IC based on subsampled data, we need to approximate two parts of IC, $\log p(\boldsymbol{x} \mid \boldsymbol{x})$ for goodness of fit, and $\text{tr}\left[J^{-1}(\widehat{\theta})I(\widehat{\theta})\right]$ for model complexity penalty, where I and J are defined in Equations (9) and (10).

Let $\boldsymbol{x} = (x_1, \cdots, x_n)$ be n independent observations from the cumulative distribution function $F(\widetilde{x})$ with probability density function $p(\widetilde{x})$, $\eta_1, \cdots, \eta_n$ be sampling probabilities, respectively, and θ be the unknown parameter for $p(\widetilde{x})$. The subsampled IC is defined as

$$\text{IC}_{\text{sp}} = -2\log\left[\frac{1}{B}\sum_{b=1}^{B} \exp\{n\ell_r(\theta^b)\}\right] + n \cdot \text{tr}\left[J_{\text{sp}}^{-1}(\widehat{\theta})I_{\text{sp}}(\widehat{\theta})\right]. \tag{13}$$

For the b-th MCMC iteration, $\ell_r(\theta^b) = \frac{\log\{p(y_i^*|\theta^b)\}}{n\eta_i^*}/r$, where $x_1^*, \cdots, x_r^*$ are r subsampled observations with respective subsample probabilities $\eta_1^*, \cdots, \eta_r^*$, the expectation $\text{E}_{\theta|\boldsymbol{x}}(\ell_r(\theta))$ is estimated by

$$\text{E}_{\theta|\boldsymbol{x}}\left[\exp(\ell_r(\theta))\right] = \frac{1}{B}\sum_{b=1}^{B} \exp\Big(\ell_r(\theta^b)\Big). \tag{14}$$

Remark 2. To approximate the penalty term $\text{tr}\left[J^{-1}(\widehat{\theta})I(\widehat{\theta})\right]$, we define

$$I_{\text{sp}}(\theta) = \frac{1}{r}\sum_{i=1}^{r} \frac{\partial\phi_{\text{sp}}(x_i^*, \theta)}{\partial\theta}\frac{\partial\phi_{\text{sp}}(x_i^*, \theta)}{\partial\theta^\top}, \tag{15}$$

and

$$J_{\text{sp}}(\theta) = -\frac{1}{r}\sum_{i=1}^{r}\frac{\partial^2\phi_{\text{sp}}(x_i^*,\theta)}{\partial\theta\partial\theta^T}, \tag{16}$$

where $\phi_{\text{sp}}(x_i^*,\theta) = \frac{1}{n\eta_i^*}\log\{p(x_i^* \mid \theta)\} + \log\{\pi_0(\theta)/(2r)\}$, and $\pi_0(\theta)$ is the prior distribution of θ. Suppose $x_1^*,\cdots,x_r^*$ are r subsampled observations with respective subsample probabilities $\eta_1^*,\cdots,\eta_r^*$, and $\widehat{\theta}$ is the posterior mean obtained via subsampled MCMC, the penalty term is therefore approximated by based on results in Ai et al. (2018):

$$\operatorname{tr}\left[J^{-1}(\widehat{\theta})I(\widehat{\theta})\right] \approx n \times \operatorname{tr}\left[J_{\text{sp}}^{-1}(\widehat{\theta})I_{\text{sp}}(\widehat{\theta})\right]. \tag{17}$$

The detailed technical procedures to calculate $I_{\text{sp}}(\theta)$ and $J_{\text{sp}}(\theta)$ in linear regression and logistic regression are given in the Appendix.

3.3 Properties of the Subsampled Information Criteria

Let $\eta_1,...,\eta_n$ be nonuniform subsampling probabilities such that $0 < \eta_i < 1,\ i = 1,...,n$ and $\sum_{i=1}^{n}\eta_i = 1$. For a subsample, $x_1^*,...,x_r^*$ taken randomly according to η_i's with replacement and $\ell(x_i \mid \theta)$ is the log probability density given parameter θ, we have

$$\mathrm{E}\left[\frac{1}{r}\sum_{i=1}^{r}\frac{\ell(x_i^* \mid \theta)}{\eta_i^*}\right] = \sum_{i=1}^{n}\ell(x_i \mid \theta). \tag{18}$$

Let $\mathcal{G} = \{g(\widetilde{x} \mid \theta)\}$ be a family of candidate statistical models. The quantity $\nu = \mathrm{E}_{\widetilde{x}}[\mathrm{E}_{\theta|\widetilde{x}}[\log g(\widetilde{x} \mid \theta)]]$ is to measure the deviation of the approximating model from the true model. For DIC, the estimator for ν is $\widehat{\nu} = \frac{1}{n}\mathrm{E}_{\theta|\boldsymbol{x}}[\log L(\theta \mid \boldsymbol{x})]$. From Ando (2007), $\widehat{\nu}$ is generally positively biased as an estimator of ν. The bias correction term is defined as

$$b_\nu = \mathrm{E}_{\boldsymbol{x}}\left[\widehat{\nu} - \nu\right]. \tag{19}$$

The bias correction term has the same definition in Ando (2007). Therefore the bias correct of posterior mean of log likelihood is

$$\mathrm{E}_{\theta|\boldsymbol{x}}[\log L(\boldsymbol{x} \mid \theta)]/n - \widehat{b}_\nu \approx \mathrm{E}_{\theta|\boldsymbol{x}}[l_r(\theta)]/n - \widehat{b}_\nu,$$

where $\widehat{b}_\nu$ is the estimated bias correction term. Under the framework of Spiegelhalter et al. (2002), we can plug in the unbiased estimator of log likelihood using subsampled data, and obtain

$$\mathrm{DIC}_{\text{sp}} = -2\mathrm{E}_{\theta|\boldsymbol{x}}[\ell_r(\theta)] + p_D, \tag{20}$$

where $p_D = 2[\log L(\boldsymbol{x} \mid \widehat{\theta}) - E_{\theta|\boldsymbol{x}}[\ell_r(\theta)]]$, $\widehat{\theta}$ being the posterior mean. From direct calculation, we have

$$\mathrm{E}[\mathrm{DIC}_{\text{sp}} \mid \widehat{\theta}] = \mathrm{DIC}, \tag{21}$$

where the expectation is taken with respect to the randomness of subsampling only. Similarly from (18) , we have based on results in Ai et al. (2018):

$$\mathrm{E}\left[\log\left\{\frac{1}{B}\sum_{b=1}^{B}\exp\Big(\ell_r(\theta^b)\Big)\right\}\right] = \log\left\{\frac{1}{B}\sum_{b=1}^{B}\exp\Big(l_r(\theta^b)\Big)\right\},$$

$$\mathrm{E}\left[\operatorname{tr}\left[J_{\text{sp}}^{-1}(\widehat{\theta})I_{\text{sp}}(\widehat{\theta})\right] \mid \widehat{\theta}\right] = \frac{1}{n}\operatorname{tr}\left[J^{-1}(\widehat{\theta})I(\widehat{\theta})\right].$$

Furthermore, we have

$$\mathrm{E}[\mathrm{IC}_{\text{sp}} \mid \widehat{\theta}] = \mathrm{IC}, \tag{22}$$

where the expectation is taken with respect to the randomness of subsampling only.

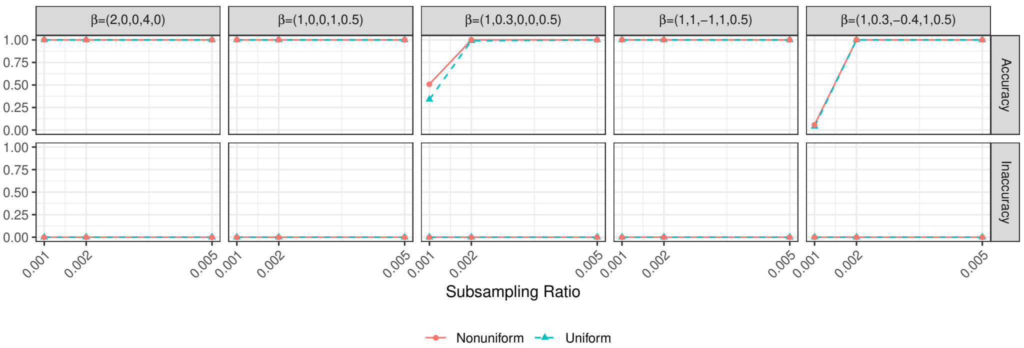

Figure 1: Selection accuracy (inaccuracy) by DIC$_{\text{sp}}$for different subsampling ratios in linear regression.

4. Simulation Studies

In this section, we evaluate the performance of the proposed criteria under three different designs: linear regression, balanced-outcome logistic regression, and imbalanced-outcome logistic regression. Comparison with uniform subsampling procedures for different subsampling ratios is also performed.

4.1 Bayesian Linear Regression Model Selection

We use a setting similar to that in Shao (1997). The size of the full data in each replicate is set to $n = 100,000$. For each observation i, five covariates x_{i1}, x_{i2}, . . ., x_{i5} are generated i.i.d. from $N(0, 1)$, and the dependent variable is generated such that

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_5 x_{i5} + \epsilon_i, \quad i = 1, \ldots, n, \tag{23}$$

where ϵ_i are again i.i.d. $N(0, 1)$. We consider 5 true parameter vectors. The MCMC chain length is set to be 15,000 with the first 5,000 as burn-in. We consider three different subsampling ratios, namely, r/n. A total of 100 replicates are performed. The selection results based on DIC$_{\text{sp}}$ and IC$_{\text{sp}}$ are presented in Figures 1, 2. The accuracy is defined as the proportion of replicates where the optimal correct model is selected out of the 100 replicates. In addition, the inaccuracy, defined as the proportion of replicates where incorrect models are selected, is reported in the parentheses. For example, if the true model is $\boldsymbol{\beta} = (2, 0, 0, 4, 0)$, then the model only selecting X_1 and X_4 is the optimal correct model, the models containing X_1 and X_4 are correct models, and the rest models are incorrect models. From Figure 1 we observe that, the DIC$_{\text{sp}}$ based on both uniform and nonuniform sampling schemes perform very well, giving 0 inaccuracy under all scenarios and sampling ratios and accuracy equal to 1 under three scenarios. Even if the accuracies of two scenarios at 0.001 sampling ratio is not 1, it increases to 1 at sampling ratio 0.002. It is noticed from Figure 2 that the IC$_{\text{sp}}$ has rather low accuracy when the true model is small, i.e. there are many zero-effect covariates, for both uniform and nonuniform subsample schemes, but meanwhile, the inaccuracy are basically 0 under all scenarios and sampling ratios, which means that correct models which contain the true covariates but are larger than the optimal correct model are highly selected. This indicates that, compared to DIC$_{\text{sp}}$, IC$_{\text{sp}}$ is relatively a conservative procedure, and tends not to err by incorporating more than just necessary covariates. One reason for IC$_{\text{sp}}$ tending to select larger models than DIC$_{\text{sp}}$ is that IC has less penalty than DIC on model complexity. Another partial reason is that our nonuniform

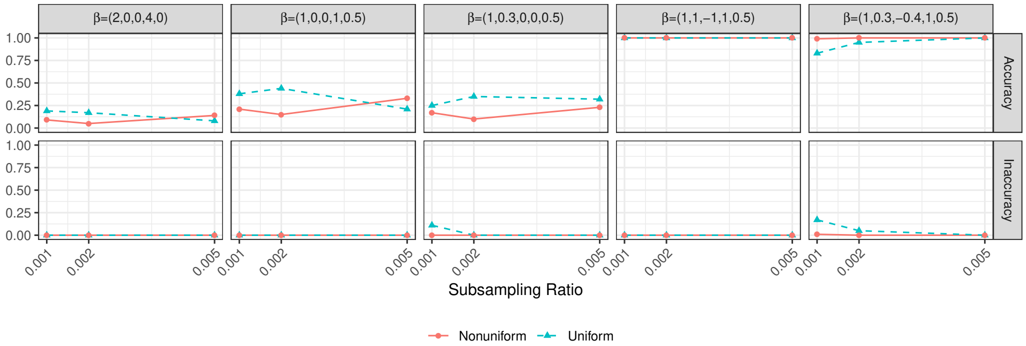

Figure 2: Selection accuracy (inaccuracy) by IC_{sp} for different subsampling ratios in linear regression.

sampling procedure places more emphasis on approximating the log-likelihood, which will tend to select more complex models to increase goodness-of-fit.

4.2 Bayesian Logistic Regression Model Selection

We use the logistic regression to demonstrate the usage of our proposed procedures for generalized linear models. Both balanced and imbalanced designs are considered. For the balanced design, five covariates are generated i.i.d. from $N(0, 1)$. For the imbalanced design, in addition to a constant term of $x_{i1} \equiv 1$, four covariates are generated i.i.d. from $\text{Uniform}(0, 1)$. Each y_i is generated from $\text{Bernoulli}(1/1 + \exp\big(-\boldsymbol{x}_i^\top \boldsymbol{\beta}\big))$. A total of 100 replicates are ran for each model and for each subsampling ratio, 0.001, 0.002, and 0.005. The MCMC chain length is 15,000 with the first 5,000 as burn-in. The results of accuracies and inaccuracies are presented in Figures 3, 4, 5 and 6.

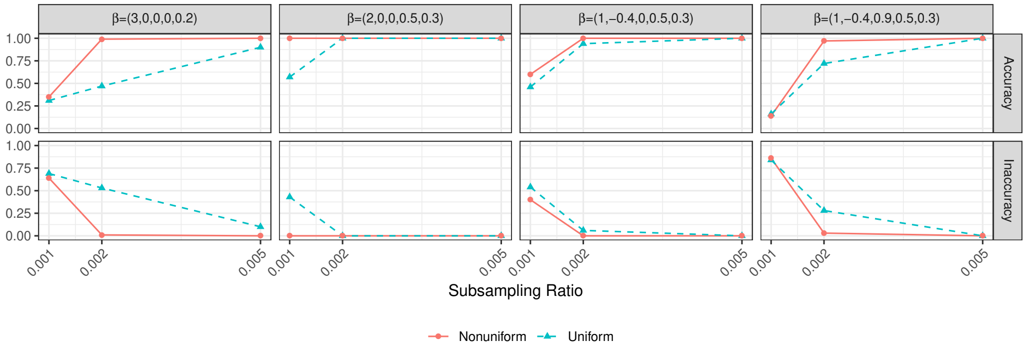

Figure 3: Selection accuracy (inaccuracy) by DIC_{sp} for different subsampling ratios in logistic regression when outcomes are balanced.

It is seen from the figures that, for both designs and for both DIC_{sp} and IC_{sp}, under almost all scenarios and subsampling ratios, the proposed nonuniform subsampling scheme performs significantly better than the uniformly subsampling scheme in terms of higher accuracies and lower inaccuracies. This difference becomes less significant with the increase in the subsampling ratio, r/n, but for some models it is still quite large even when $r/n = 0.005$. It is also observed that, the selection accuracies of IC_{sp}, though still increasing as subsample sizes increase, are lower than the accuracies of DIC_{sp}, which means that IC_{sp} requires more subsamples than DIC_{sp} to select true models. In addition, we notice that

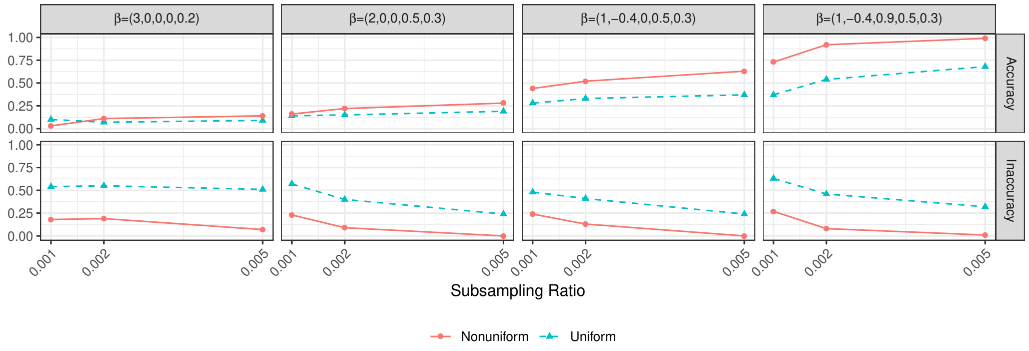

Figure 4: Selection accuracy (inaccuracy) by $\mathrm{IC}_{\mathrm{sp}}$ for different subsampling ratios in logistic regression when outcomes are balanced.

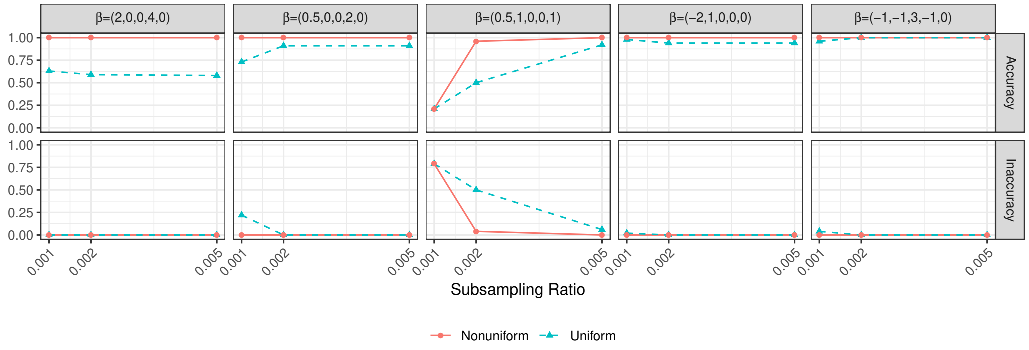

Figure 5: Selection accuracy (inaccuracy) by $\mathrm{DIC}_{\mathrm{sp}}$ for different subsampling ratios in logistic regression when outcomes are imbalanced.

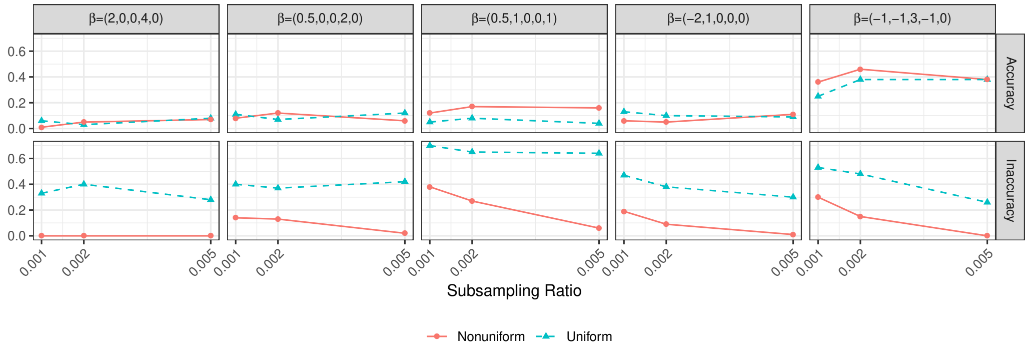

Figure 6: Selection accuracy (inaccuracy) by $\mathrm{IC}_{\mathrm{sp}}$ for different subsampling ratios in logistic regression when outcomes are imbalanced.

Table 1: Model selection results of Census Income data using subsampling schemes

Subsample Size	Criterion	Uniform		NonUniform	
		Selected Covariates	AUC	Selected Covariates	AUC
$r = 200$	DIC$_{\text{sp}}$	1, 2, 4, 5, 6	0.8153	1, 2, 4, 5, 6	0.8153
	IC$_{\text{sp}}$	1, 2, 4, 5, 6	0.8153	1, 2, 3, 4, 5, 6	0.8155
$r = 500$	DIC$_{\text{sp}}$	1, 2, 4, 5, 6	0.8152	1, 2, 4, 5, 6	0.8153
	IC$_{\text{sp}}$	1, 2, 4, 5, 6	0.8152	1, 2, 3, 4, 5, 6	0.8154
$r = 1000$	DIC$_{\text{sp}}$	1, 2, 4, 5, 6	0.8153	1, 2, 4, 5, 6	0.8152
	IC$_{\text{sp}}$	1, 2, 3, 4, 5, 6	0.8153	1, 2, 3, 4, 5, 6	0.8154

IC$_{\text{sp}}$ gives similar inaccuracies across scenarios, but reaches much higher accuracies when the true models are larger, i.e. when the true models contain less zero-effect covariates, in other words, IC$_{\text{sp}}$ tends to select larger correct models than DIC$_{\text{sp}}$, which is similar to the case in linear regression. One possible remedy is to change the subsampling strategy based on the posterior predictive distribution.

5. Application to Real Data

5.1 Census Income Data Set

In this section, we apply proposed model selection methods to a census income dataset Kohavi (1996) extracted from the 1994 Census database. The response is whether a person's income exceeds $50,000 a year. There are in total 48,842 observations in this dataset, and 11,687 of them have income over $50,000 a year. We consider 5 covariates together with the intercept in our analysis. The covariates used are Age of individuals, Final weight, Gender indicator, Highest level of education in numerical form, Hours worked per week, and we denote the intercept and these covariates by X_1, X_2, X_3, X_4, X_5 and X_6. The variable Final weight (X_3) means the number of people the individual represents. Gender indicator (X_4) is a binary variable, and the rest are continuous variables. We first conduct a logistic regression on this dataset, and all covariates are significant under significance level of 0.05. Choosing non-informative priors for regression coefficients, we run 50,000 MH iterations and drop the first 15,000 as burn-in to perform both uniform and nonuniform subsampling schemes to the dataset, subsample sizes chosen as $r = 200,\ 500$, and 1000. We select the optimal correct models based on DIC$_{\text{sp}}$ and IC$_{\text{sp}}$, and use the estimated coefficients to calculate the area under curve (AUC; Zhou et al., 2009) of selected models. The results are shown in Table 1. We also run Bayesian process under the same setting using full data, by which all covariates are selected by both DIC and IC, and the AUC under full model is 0.8154.

From Table 1 we see that IC$_{\text{sp}}$ tends to select larger models than DIC$_{\text{sp}}$, which is similar to the conclusion we have in simulation studies. In addition, overall speaking, the DIC$_{\text{sp}}$ and IC$_{\text{sp}}$ based on the nonuniform subsampling scheme outperform the DIC$_{\text{sp}}$ and IC$_{\text{sp}}$ uniform subsampling scheme, for selecting more significant covariates and obtaining higher AUC values.

5.2 Forest Cover Type Data

In order to assess the performance of the proposed methods, we consider the dataset named "Covertype" from the UCI Machine Learning Repository (Dheeru and Karra Taniskidou, 2017). The dataset contains 581,012 observations, and 54 covariates, 10 of which being

Table 2: Model selection results of Covertype data using subsampling schemes

Subsample Size	Criterion	Uniform		NonUniform	
		Selected Covariates	AUC	Selected Covariates	AUC
$r = 500$	DIC$_{\text{sp}}$	1, 2, 3, 4, 6, 8, 11	0.7734	1, 2, 3, 4, 6, 7, 8, 11	0.7792
	IC$_{\text{sp}}$	1, 2, 4, 6, 7, 8	0.7673	1, 2, 3, 4, 5, 6, 7, 8, 11	0.7789
$r = 1000$	DIC$_{\text{sp}}$	1, 2, 3, 4, 6, 7, 10, 11	0.7774	1, 2, 3, 4, 6, 7, 8, 11	0.7791
	IC$_{\text{sp}}$	1, 2, 4, 8, 9, 11	0.7630	1, 2, 3, 4, 6, 7, 9, 10, 11	0.7793

continuous, and the rest being binary. We consider 10 continuous covariates together with the intercept in our analysis. Similar to in Collobert et al. (2002), we combine 6 classes and modify the multi-class problem into an imbalanced binary classification problem, with 14.8% 1's and 85.2% 0's. As a benchmark, we first conduct the stepwise logistic regression based on the Akaike information criterion (AIC; Akaike, 1973) with the entire dataset. The selected model contains X_1(Intercept), X_2, X_3, X_4, X_5, X_6, X_7, X_8 and X_{11}, and the corresponding AUC is 0.7790. We then apply uniform and nonuniform subsampling schemes with subsample sizes r =500 and 1000. In each case, 100,000 MH iterations are ran, with the first 2,000 samples thrown as burn-in, and we store one sample for every 20 MH iterations in order to reduce autocorrelations of MH samples. The results are reported in Table 2.

It is observed that, under both subsample sizes and both selection criterion, the final models selected by the proposed nonuniform subsampling scheme are closer to the step regression result and achieve higher AUC values than the models selected by uniform subsampling. It is also noteworthy that, the proposed nonuniform subsampling achieves better selection results and higher AUC values at $r = 500$ than uniform subsampling at $r = 1000$. Again, under nonuniform subsampling scheme, IC$_{\text{sp}}$ tends to choose larger models than DIC$_{\text{sp}}$, and IC$_{\text{sp}}$ needs more subsamples under uniform subsampling scheme to obtain more significant results than nonuniform subsampling scheme, which is similar to the results obtained in simulation studies.

6. Discussion

In this paper, we proposed the subsampled DIC and IC which approximate the full data DIC and IC for big data based on subsampled MCMC. Our subsampled approaches overcome the computation bottleneck of big data. In the simulation study, we found that nonuniform subsampling strategy performs better than the uniform subsampling strategy in small subsample size cases and in rare events data case of logistic regression.

In addition, four topics beyond the scope of this paper are worth further investigation. First, we need to theoretically justify the model selection powers of the subsampled DIC and IC based on the KL divergence. Second, all subsample sizes in this work are prespecified. In the future, an adaptive subsampling scheme that automatically calculates the number of subsampled observations needed is desirable. A subsampling strategy that is based on the posterior predictive distribution is also devoted to future research. Our proposed model selection criteria are mainly based on algorithm 2. Modifying our proposed criteria in other subsampling algorithm in Bardenet et al. (2014); Quiroz et al. (2018) is also an important future work.

References

Ai, M., J. Yu, H. Zhang, and H. Wang (2018). Optimal subsampling algorithms for big data generalized linear models. *arXiv preprint arXiv:1806.06761*.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiado.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika 94*(2), 443–458.

Ando, T. and R. Tsay (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting 26*(4), 744–763.

Bardenet, R., A. Doucet, and C. Holmes (2014). Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *International Conference on Machine Learning*, pp. 405–413.

Bardenet, R., A. Doucet, and C. Holmes (2017). On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research 18*(1), 1515–1557.

Collobert, R., S. Bengio, and Y. Bengio (2002). A parallel mixture of SVMs for very large scale problems. In *Advances in Neural Information Processing Systems*, pp. 633–640.

Dheeru, D. and E. Karra Taniskidou (2017). UCI machine learning repository.

Hu, G. and H. Wang (2018). Minimax optimal subsampled Markov chain Monte Carlo. Technical Report 18-27, University of Connecticut, Department of Statistics.

Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *KDD*, Volume 96, pp. 202–207. Citeseer.

Korattikara, A., Y. Chen, and M. Welling (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *International Conference on Machine Learning*, pp. 181–189.

Quiroz, M., R. Kohn, M. Villani, and M.-N. Tran (2018). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association* (Forthcoming), 1–35.

Quiroz, M., M.-N. Tran, M. Villani, and R. Kohn (2018). Speeding up MCMC by delayed acceptance and data subsampling. *Journal of Computational and Graphical Statistics 27*(1), 12–22.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 221–242.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Linde (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(3), 485–493.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 583–639.

Wang, X. and D. B. Dunson (2013). Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

Zhou, S. (2011). *Bayesian model selection in terms of Kullback–Leibler discrepancy*. Ph. D. thesis, Columbia University.

Zhou, X.-H., D. K. McClish, and N. A. Obuchowski (2009). *Statistical Methods in Diagnostic Medicine*, Volume 569. John Wiley & Sons.

Appendix

A.1 Technical Details for Normal Linear Regression

A.1.1 Complete Data

For the normal linear regression case, consider a more general setting where $\boldsymbol{x}_i \in \mathcal{R}^p$, $y_i \mid \boldsymbol{x}_i \sim N(\boldsymbol{x}_i^\top \boldsymbol{\theta}, \sigma_1^2)$, $i = 1, ..., n$, and $\boldsymbol{\theta}$ follows the normal prior $\pi_0(\boldsymbol{\theta}) \sim \text{MVN}(0_{p\times 1}, \sigma_2^2 I_{p\times p})$. We have $\phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) = \ell(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) + \log[\frac{\pi_0(\boldsymbol{\theta})}{2n}]$. Then

$$\begin{aligned}\phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) &= \log \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2}{2\sigma_1^2}} + \log \frac{1}{2n(2\pi\sigma_2^2)^{\frac{p}{2}}} e^{-\frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{2\sigma_2^2}}, \\ &= c - \frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})^2}{2\sigma_1^2} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{2\sigma_2^2},\end{aligned}$$

where c is some constant that does not depend on $\boldsymbol{\theta}$. Then

$$\frac{\partial \phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})}{\sigma_1^2} - \frac{\boldsymbol{\theta}}{\sigma_2^2}, \quad \frac{\partial^2 \phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top}{\sigma_1^2} - \frac{1}{\sigma_2^2} I_{p\times p}.$$

We therefore have

$$\begin{aligned}J(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top}{\sigma_1^2} + \frac{1}{\sigma_2^2} I_{p\times p} \right), \\ I(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})}{\sigma_1^2} - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right) \left(\frac{\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta})}{\sigma_1^2} - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right)^\top .\end{aligned}$$

A.1.2 Subsampled Data

Now consider a subsample of the full dataset, where the observations are now denoted as $(\boldsymbol{x}_1^*, y_1^*), \ldots, (\boldsymbol{x}_r^*, y_r^*)$. Using similar derivations as above, an approximation of $J(\boldsymbol{\theta})$ based on this subsample is obtained as

$$J_{\text{sp}}(\boldsymbol{\theta}) = \frac{1}{nr} \sum_{i=1}^{r} \frac{\boldsymbol{x}_i^* \boldsymbol{x}_i^{*\top}}{\eta_i^* \sigma_1^2} + \frac{1}{\sigma_2^2} I_{p\times p},$$

and an approximation of $I(\boldsymbol{\theta})$ based on this subsample is

$$I_{\text{sp}}(\boldsymbol{\theta}) = \frac{1}{r^2} \sum_{i=1}^{r} \left(\frac{\boldsymbol{x}_i^*(y_i^* - \boldsymbol{x}_i^{*\top} \boldsymbol{\theta})}{n\eta_i^* \sigma_1^2} - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right) \left(\frac{\boldsymbol{x}_i^*(y_i^* - \boldsymbol{x}_i^{*\top} \boldsymbol{\theta})}{n\eta_i^* \sigma_1^2} - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right)^\top .$$

A.2 Technical Details for Logistic Regression

A.2.1 Complete Data

For the logistic regression case, consider a more general setting that $y_i \sim \text{Bernoulli}(p_i)$, where $\log \frac{p_i}{1-p_i} = \boldsymbol{x}_i^\top \boldsymbol{\theta}$, and $\boldsymbol{x}_i \in \mathcal{R}^p$, $i = 1, ..., n$. Similar to the linear regression case setting, assume $\boldsymbol{\theta}$ follows the normal prior $\pi_0(\boldsymbol{\theta}) \sim \text{MVN}(0_{p\times 1}, \sigma_2^2 I_{p\times p})$. Then

$$
\begin{aligned}
\phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) &= \ell(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) + \log[\frac{\pi_0(\boldsymbol{\theta})}{2n}] \\
&= y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} - \log\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right) + \log \frac{1}{2n(2\pi\sigma_2^2)^{\frac{p}{2}}} e^{-\frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{2\sigma_2^2}}, \\
&= y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} - \log\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right) - \frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{2\sigma_2^2} + c,
\end{aligned}
$$

where where c is some constant that does not depend on $\boldsymbol{\theta}$. Denote $p_i = \exp\{\boldsymbol{x}_i^\top \boldsymbol{\theta}\}/(1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\theta}\})$, and $w_i = p_i(1 - p_i)$, then

$$
\frac{\partial \phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{x}_i(y_i - p_i) - \frac{\boldsymbol{\theta}}{\sigma_2^2}, \quad \frac{\partial^2 \phi(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top - \frac{1}{\sigma_2^2} I_{p\times p}.
$$

Therefore we have

$$
J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top + \frac{1}{\sigma_2^2} I_{p\times p} \right),
$$

$$
I(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i(y_i - p_i) - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right) \left(\boldsymbol{x}_i(y_i - p_i) - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right)^\top.
$$

A.2.2 Subsampled Data

Consider a subsample of the full dataset $(\boldsymbol{x}_1^*, y_1^*), \ldots, (\boldsymbol{x}_r^*, y_r^*)$. According to Equations (15) and (16), the approximations of $J(\boldsymbol{\theta})$ and $I(\boldsymbol{\theta})$ based on this subsample are obtained as

$$
J_{\text{sp}}(\boldsymbol{\theta}) = \frac{1}{nr} \sum_{i=1}^{r} \frac{w_i^* \boldsymbol{x}_i^* \boldsymbol{x}_i^{*\top}}{\eta_i^*} + \frac{1}{\sigma_2^2} I_{p\times p},
$$

$$
I_{\text{sp}}(\boldsymbol{\theta}) = \frac{1}{r^2} \sum_{i=1}^{r} \left(\frac{\boldsymbol{x}_i^*(y_i^* - p_i^*)}{n\eta_i^*} - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right) \left(\frac{\boldsymbol{x}_i^*(y_i^* - p_i^*)}{n\eta_i^*} - \frac{\boldsymbol{\theta}}{\sigma_2^2} \right)^\top.
$$