ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# A novel variational Bayesian method for variable selection in logistic regression models

Chun-Xia Zhang *, Shuang Xu, Jiang-She Zhang

*School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

ARTICLE INFO

*Article history:*
Received 22 January 2018
Received in revised form 21 August 2018
Accepted 28 August 2018
Available online xxxx

*Keywords:*
Variable selection
Logistic regression
Sparse model
Variational Bayes
Indicator model
High-dimensional data

ABSTRACT

With high-dimensional data emerging in various domains, sparse logistic regression models have gained much interest of researchers. Variable selection plays a key role in both improving the prediction accuracy and enhancing the interpretability of built models. Bayesian variable selection approaches enjoy many advantages such as high selection accuracy, easily incorporating many kinds of prior knowledge and so on. Because Bayesian methods generally make inference from the posterior distribution with Markov Chain Monte Carlo (MCMC) techniques, however, they become intractable in high-dimensional situations due to the large searching space. To address this issue, a novel variational Bayesian method for variable selection in high-dimensional logistic regression models is presented. The proposed method is based on the indicator model in which each covariate is equipped with a binary latent variable indicating whether it is important. The Bernoulli-type prior is adopted for the latent indicator variable. As for the specification of the hyperparameter in the Bernoulli prior, we provide two schemes to determine its optimal value so that the novel model can achieve sparsity adaptively. To identify important variables and make predictions, one efficient variational Bayesian approach is employed to make inference from the posterior distribution. The experiments conducted with both synthetic and some publicly available data show that the new method outperforms or is very competitive with some other popular counterparts.



## 1. Introduction

Logistic regression (Hosmer Jr. et al., 2013), one of the most popular tools to solve classification tasks, has always received great interest of both statisticians and machine learning researchers. Given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{+1, -1\}$, a logistic regression model hypothesizes that

$$p(y_i = 1|\mathbf{x}_i) = \sigma(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i), \tag{1}$$

where $\sigma(x)$ denotes $1/(1 + \exp(-x))$. According to maximum likelihood estimation principle, one can get an estimate of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}$, via maximizing $\prod_{i=1}^n p(y_i|\mathbf{x}_i)$. Although there is no closed-form solution of $\hat{\boldsymbol{\beta}}$, we can approximate it by the iteratively re-weighted least squares (IRLS) algorithm (Bishop, 2006). In high-dimensional situations, especially when $p \gg n$, the estimation accuracy of $\hat{\boldsymbol{\beta}}$ will be unsatisfactory because of the limited information provided by training data. Fortunately, the true model is often *sparse* in the sense that only a few covariates are truly influential to the response. By detecting these important variables and fitting a model with only these variables, both the estimation accuracy of $\hat{\boldsymbol{\beta}}$ and the prediction

* Corresponding author.
*E-mail address:* cxzhang@mail.xjtu.edu.cn (C.-X. Zhang).

https://doi.org/10.1016/j.csda.2018.08.025
0167-9473/

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

accuracy of the model can be significantly enhanced. In some applications, on the other hand, researchers may be more interested in identifying important variables so that the relationship between covariates and our interested outcome can be more easily revealed. Therefore, it is particularly important to perform variable selection efficiently and accurately.

In linear regression models, it is well-known that some coefficients can be estimated exactly as zero by shrinkage techniques such as least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), smoothly clipped absolute deviation penalty (SCAD) (Fan and Li, 2001), minimax concave penalty (MCP) (Zhang, 2010) as well as $L_{1/2}$-norm penalty (Xu et al., 2012). These techniques can be naturally applied to handle variable selection problems in logistic regression by means of penalized IRLS algorithms. A representative but incomplete list of references includes Breheny and Huang (2011), Jiang and Zhang (2014), Krishnapuram et al. (2005) Liang et al. (2013) and Tian et al. (2008). It is worthwhile that shrinkage methods demand careful selection of their associated tuning parameters. In high-dimensional situations, it is often challenging to choose an appropriate tuning parameter so that truly important variables can be exactly detected.

Moreover, one also can infer from the posterior distribution of $\boldsymbol{\beta}$ in (1) from the Bayesian viewpoint. Note that there is no conjugate prior in exponential family of distributions for logistic regression. A common trick is to approximate likelihood by a quadratic function, such as Gaussian approximation (Spiegelhalter and Lauritzen, 1990; Mackay, 1992) or local variational Bayesian methods (Jaakkola and Jordan, 2000). Thereafter, the Gaussian conjugate prior becomes available. Recently, Polson et al. (2013) developed an exact and simple Bayesian logit regression model to make predictions by introducing a Pólya-Gamma latent variable. When processing high-dimensional data with Bayesian methodologies, the *indicator model* is a popular way to identify important variables (Latouche et al., 2016; O'Hara and Sillanpää, 2009; Ormerod et al., 2017; Park and Casella, 2008; Ročková and George, 2014). As a matter of fact, the indicator model equips each covariate $X_j$ with a binary latent variable $\gamma_j$, where $\gamma_j \in \{0, 1\}$ follows a Bernoulli distribution (Kuo and Mallick, 1998). The covariate $X_j$ is important if $\gamma_j = 1$ and unimportant otherwise. In practice, the variables with posterior probability of $\gamma_j = 1$ being less than 0.5 are deemed as unimportant. To infer from the posterior distribution of unknown parameters such as $\gamma_j$ and $\beta_j$, Markov Chain Monte Carlo (MCMC) algorithms are widely used (Kyung et al., 2010; Ghosh et al., 2018). Moreover, Nott and Leonte (2004) considered an indicator model for generalized linear regression and they employed a variant of MCMC strategy based on the Swendsen–Wang sampling algorithm to draw samples. Tüchler (2008) used the auxiliary mixture sampling to achieve both variable and covariance selection in the context of logistic mixed effects models. By imposing non-local prior densities (Rossell and Rubio, 2018; Rossell and Telesca, 2017) on regression coefficients, Nikooienejad et al. (2016) proposed a Bayesian variable selection method called iMOMLogit for binary outcomes. To sample from posterior distribution with an MCMC sampler, iMOMLogit utilizes a Laplace approximation to alleviate the computational burden. Unfortunately, it has been shown that MCMC is too computationally intensive when the number of covariates $p$ is large (Blei et al., 2017).

Therefore, many researchers attempt to design an alternative of MCMC-based methods. According to the mechanism of empirical Bayes, Pungpapong et al. (2015) proposed an iterated conditional modes/medians (ICMM) algorithm to select informative variables from massive candidates in linear regression models. At present, the ICMM algorithm has been extended to the situations of generalized linear regression models (Pungpapong et al., 2017). It assumes that each regression coefficient $\beta_j$ ($j = 1, \ldots, p$) is governed by a spike and slab mixture prior, that is,

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\delta_0(\beta_j) + \gamma_j f(\beta_j), \tag{2}$$

where $\delta_0(\cdot)$ is the Dirac delta function at zero and $f(\cdot)$ is a probability density function (pdf). Hence, $\beta_j$ is zero if $\gamma_j = 0$, otherwise it is drawn from $f(\beta_j)$. To further enhance sparsity, ICMM assumes that $f(\cdot)$ is the pdf of a Laplace distribution, which in fact corresponds to an $L_1$-norm penalty. The parameters $\gamma_j$ and $\beta_j$ are updated according to conditional modes and posterior medians, respectively. This model works well and usually converges within a few steps as long as good initial values are provided. On the other hand, however, ICMM tends to generate many false negatives. Actually, the model considered by Carbonetto and Stephens (2012) is similar to ICMM, with the exception that $f(\cdot)$ comes from Gaussian distribution family. In addition, they inferred from the posterior distribution via a variational Bayesian method and then approximated the posterior of $\gamma_j = 1$ by means of importance sampling. On account of $\gamma_j$ being a latent variable, the expectation–maximization (EM) algorithm is very suitable for indicator models. Ročková and George (2014) developed an EM-based variable selection (EMVS) algorithm for identifying important variables in linear regression models. At present, the EMVS algorithm has been generalized to logistic regression models (Koslovsky et al., 2018; Mcdermott et al., 2016). In particular, EMVS imposes a spike and slab Gaussian mixture prior for each regression coefficient $\beta_j$. With respect to each indicator variable $\gamma_j$ ($j = 1, \ldots, p$), it assigns independent and identically distributed (i.i.d.) Bernoulli priors to them, i.e., $\gamma_j \mid \rho \sim \text{Ber}(\rho)$. In addition, a beta hyper-prior distribution is employed for $\rho$. Despite EMVS alleviates the high computational cost of MCMC sampling to a certain degree, sometimes it may be slow due to intrinsic features of EM algorithm. Besides this, the EM algorithm may often be trapped into a local minimum. Ročková and George (2014) suggested to solve this by applying EM algorithms several times in which the estimate obtained in last iteration is taken as the initial value in next iteration. Recently, some other variants of EM algorithms for variable selection in linear regression have been designed to avoid the local minimum (Wang et al., 2016; Ročková, 2017). More importantly, different from linear regression, the crux of logistic regression lies in that there is no closed-form solution in the M-step and the Newton–Raphson algorithm can be utilized. Since some iterative method needs to be used to find the corresponding solution, it is often time-consuming to apply EMVS in logistic regression (Koslovsky et al., 2018; Mcdermott et al., 2016).

As an alternative of variational Bayes, the expectation propagation (EP) algorithm is also an efficient tool to approximate posterior distributions with relatively low computational burden. In this aspect, Hernández-Lobato et al. (2015) developed

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

a novel EP algorithm to implement approximate inference in linear regression models with spike and slab priors. Except for the assumed different models, their purpose is to achieve high prediction accuracy instead of accurate identification of important variables. Furthermore, Horii (2017) discussed the application of variational Bayesian method in sparse logistic regression by assuming hierarchical priors for unknown parameters. Wang and Blei (2013) brought forward two variational inference methods (i.e., Laplace VI and delta method VI) for nonconjugate models in which Bayesian logistic regression is a special case. To the best of our knowledge, their modeling goal is to make good predictions instead of detecting important variables for better interpretability.

Motivated by the variational Bayesian method put forward by Ormerod et al. (2017) to perform variable selection in linear regression models, we propose in this paper a Bayesian indicator model for logistic regression to identify important variables. In the novel model, all $\gamma_j$ and $\beta_j$ are assumed to be independent. The indicator $\gamma_j$ is governed by the Bernoulli distribution $\mathrm{Ber}(\rho)$ and each coefficient $\beta_j$ is assigned with a separate shrinkage prior. As shown in later experiments (Section 3), this type of prior enjoys the advantage to avoid the preference of the null model. The posterior distributions of all parameters are approximated by a variational Bayesian approach. In addition, the sparsity of the model is controlled by the hyperparameter $\rho$. To select an optimal value for $\rho$, two schemes are proposed. One is to use grid search to minimize Bayesian information criterion (BIC), while the other is to impose a Beta distribution on $\rho$ and then employ the variational method to infer its value. The experiments conducted with both synthetic and some publicly available data demonstrate that our model outperforms or is very competitive with several other popular alternatives.

The remainder of the paper is organized as follows. Section 2 describes all the details about our proposed variational Bayesian variable selection model for logistic regression. Particularly, the formulation of the model, how to make variational inference of unknown (hyper)parameters, the initialization of (hyper)parameters as well as the method to predict new data are presented in the corresponding subsections of Section 2. In Sections 3 and 4, experiments are conducted with synthetic and some publicly available data to examine and compare the performance of the proposed method with several other popular techniques, respectively. Section 5 discusses the difference between our method and some other closely related counterparts. Finally, Section 6 includes the conclusions of the paper. In appendix, we present the detailed process to make inference of unknown (hyper)parameters with a variational Bayesian technique.

## 2. Variational Bayesian variable selection for logistic regression

### 2.1. Model formulation

To ease expositions, we use bold uppercase and lowercase letters to denote matrices and vectors, respectively. The scalars are referred to as non-bold letters. In later discussions, $\mathbf{X}_j$ indicates the $j$th column of the matrix $\mathbf{X}$ while $\mathbf{x}_i^{\mathrm{T}}$ means the $i$th row of $\mathbf{X}$. The superscript T represents the transpose of a matrix or a vector. Meanwhile, $x_{ij}$ denotes the $(i, j)$th entry of $\mathbf{X}$. The symbol $\odot$ stands for the Hadamard product. Given a matrix $\mathbf{A}$, we use the notations $\det(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$ to denote its determinant and trace, respectively.

As presented in introduction, the logistic regression generally assumes that the response variable $Y \in \{-1, 1\}$ follows a Bernoulli distribution. Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, there is $p(y_i) = 1/\left(1 + \exp\left(-y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}\right)\right) = \sigma(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})$ and $\boldsymbol{\beta}$ denotes an unknown coefficient vector. For the purpose of selecting important predictors, we further equip $\beta_j$ with a binary latent variable $\gamma_j$, where $\gamma_j = 1$ indicates that the $j$th predictor is important and $\gamma_j = 0$ means that it is unimportant. With the above assumptions, there is $p(y_i) = \sigma(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Gamma} \boldsymbol{\beta})$, where $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$.

By reformulating the $n$ observations as $\mathbf{X} = (\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$, the likelihood can be expressed as

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\Gamma}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \sigma(y_i \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\Gamma} \boldsymbol{\beta}), \tag{3}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ denotes the $i$th observation. Unlike the situation in linear regression models, there is no conjugate prior for $\boldsymbol{\beta}$ in this model. In what follows, an approximation technique will be used to convert the likelihood in (3) into a quadratic function of $\boldsymbol{\beta}$. Here, we consider the Gaussian prior for $\boldsymbol{\beta}$, namely,

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\beta}|\mathbf{0}, \mathbf{A}^{-1}) = \frac{(\det(\mathbf{A}))^{1/2}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}} \mathbf{A} \boldsymbol{\beta}\right), \tag{4}$$

where $\det(\mathbf{A})$ denotes the determinant of the precision matrix $\mathbf{A}$. To simplify discussions, we suppose $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha}) = \mathrm{diag}(\alpha_1, \alpha_2, \ldots, \alpha_p)$ in which $\alpha_1, \ldots, \alpha_p$ are hyperparameters. For $\boldsymbol{\alpha}$, we assume its elements be independent. And a conjugate hyper-prior modeled by a Gamma distribution is imposed on each $\alpha_j$, that is,

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{p} \mathrm{Gam}(\alpha_j|a_0, b_0), \tag{5}$$

in which $a_0$, $b_0$ are shape and scale parameters, respectively. As for the latent vector $\boldsymbol{\gamma}$, we assume that

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} \mathrm{Ber}(\gamma_j|\rho), \tag{6}$$

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

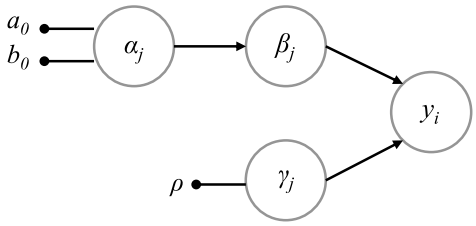



**Fig. 1.** The graphical representation of the Bayesian logistic regression model. Here, $\rho$, $a_0$ and $b_0$ are hyperparameters.

by following the common practice (Ormerod et al., 2017; Ročková and George, 2014), in which $\text{Ber}(\cdot|\rho)$ indicates the Bernoulli distribution with the success rate parameter $\rho$. Here, it is worth pointing out that the hyperparameter $\rho$ in (6) plays an important role in controlling sparsity. To facilitate the understanding of model configurations, Fig. 1 depicts the graphical representation of the above Bayesian logistic regression. In next subsection, we will show how to make inference of all parameters and hyperparameters with a variational Bayesian method.

### 2.2. Variational inference

To infer from the posterior distribution of parameters, we employ a variational Bayesian method to approximate their posterior distributions. The basic idea is to minimize the *Kullback–Leibler (KL) divergence* between a variational distribution $q(\mathbf{z})$ and the true posterior distribution $p(\mathbf{z}|\mathcal{D})$, where $q(\mathbf{z}) = \prod_{i=1}^{k} q(z_i)$ and $z_1, \ldots, z_k$ are variational parameters, $\mathcal{D}$ means the given training data. In theory, the minimization of KL divergence is equivalent to maximizing the evidence lower bound (ELBO) defined as

$$\mathcal{L}(q) = \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{p(\mathbf{z}, \mathcal{D})}{q(\mathbf{z})}\right] \leq \log p(\mathcal{D}), \tag{7}$$

where $p(\mathcal{D})$ is the model evidence. According to the main principle of variational inference, it has been shown (Blei et al., 2017) that the optimal solution satisfies

$$q(z_i) \propto \exp\left\{\mathbb{E}_{-q(z_i)} \log p(\mathbf{z}, \mathcal{D})\right\}, \tag{8}$$

where $\mathbb{E}_{-q(z_i)}$ denotes the expectation with respect to all parameters but $z_i$. By iteratively updating variational distribution $q(z_1), q(z_2), \ldots, q(z_k)$, the ELBO in (7) can be made to monotonically increase.

With the model assumptions in Section 2.1, our Bayesian model can be formulated as

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma} &\sim \text{Ber}\left(\mathbf{y} \mid \sigma(\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta})\right), \\ \boldsymbol{\beta} \mid \boldsymbol{\alpha} &\sim \mathcal{N}(\boldsymbol{\beta} \mid \mathbf{0}, \mathbf{A}^{-1}), \\ \boldsymbol{\alpha} &\sim \prod_j \text{Gam}(\alpha_j|a_0, b_0), \\ \boldsymbol{\gamma} &\sim \prod_j \text{Ber}(\gamma_j|\rho). \end{aligned} \tag{9}$$

Therefore, the joint distribution of all parameters (i.e., $p(\mathbf{z}, \mathcal{D})$ in ELBO as defined in (7)) is

$$p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\boldsymbol{\beta}|\boldsymbol{\alpha})p(\boldsymbol{\gamma})p(\boldsymbol{\alpha}). \tag{10}$$

Because of the typical form of $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\boldsymbol{\beta}$ does not have a conjugate prior. Thus, we employ a lower bound approximation for the sigmoid function (Jaakkola and Jordan, 2000), i.e.,

$$\sigma(x) \geq \sigma(\xi) \exp\left[(x - \xi)/2 - \delta(\xi)(x^2 - \xi^2)\right], \tag{11}$$

where $\delta(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]$ and $\xi$ denotes a local variational parameter. By applying (11) to $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$, we have

$$\begin{aligned} \log p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) &\geq \log h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \\ &= \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}\boldsymbol{\beta} + \sum_{i=1}^{n}\left[\log\sigma(\xi_i) - \frac{\xi_i}{2} + \delta(\xi_i)\xi_i^2\right], \end{aligned} \tag{12}$$

where $\mathbf{S} = \sum_{i=1}^{n}\delta(\xi_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$. Different from the traditional variational inference, our ELBO is

$$\widetilde{L}(q, \boldsymbol{\xi}) = \mathbb{E}_q\left[\log \frac{h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})p(\boldsymbol{\beta}|\boldsymbol{\alpha})p(\boldsymbol{\gamma})p(\boldsymbol{\alpha})}{q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})}\right]. \tag{13}$$

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

Specifically, let the variational distribution be $q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = q(\boldsymbol{\beta})q(\boldsymbol{\gamma})q(\boldsymbol{\alpha})$. Then, the variational posteriors can be evaluated by standard variational methods. With some derivations (see Appendix A for the details), we can acquire that

$$\begin{aligned} &q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ &q(\alpha_j) = \mathrm{Gam}(\alpha_j \mid a_j, b_j), \\ &q(\gamma_j) = \mathrm{Ber}(\gamma_j \mid \theta_j),\ j = 1, \ldots, p, \end{aligned} \tag{14}$$

where

$$\begin{aligned} &\boldsymbol{\Sigma} = \left[\mathrm{diag}\left(\frac{a_1}{b_1}, \ldots, \frac{a_p}{b_p}\right) + 2\mathbf{S} \odot \boldsymbol{\Omega}\right]^{-1}, \quad \boldsymbol{\Omega} = \boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}} + \boldsymbol{\Theta}(\mathbf{I}_p - \boldsymbol{\Theta}), \quad \boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta}), \\ &\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\Sigma}\boldsymbol{\Theta}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \quad \mathbf{D} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}, \\ &a_j = a_0 + \frac{1}{2}, \quad b_j = b_0 + \frac{1}{2}d_{jj}, \\ &\theta_j = \sigma\left(\frac{1}{2}\mu_j\mathbf{y}^{\mathrm{T}}\mathbf{X}_j - s_{jj}d_{jj} - 2\sum_{i\neq j} s_{ij}d_{ij}\theta_i + \log\frac{\rho}{1-\rho}\right), \end{aligned} \tag{15}$$

in which $\mathbf{I}_p$ indicates an identity matrix of order $p$, $d_{jj}$ denotes the $j$th diagonal element of $\mathbf{D}$ and $s_{jj}$ is defined similarly.

By taking expectation of each item in the logarithmic function of (13) with respect to the variational distribution, the ELBO for our model is available as

$$\begin{aligned} \widetilde{L}(q, \boldsymbol{\xi}) = &\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\log[\det(\boldsymbol{\Sigma})] + \sum_{i=1}^{n}\left[\log\sigma(\xi_i) - \frac{\xi_i}{2} + \delta(\xi_i)\xi_i^2\right] + p\left[a_0\log b_0 - \log\Gamma(a_0)\right] \\ &+ \sum_{j=1}^{p}\left[\log\Gamma(a_j) + a_j - b_0\frac{a_j}{b_j} - a_j\log b_j + \theta_j\log\frac{\rho}{\theta_j} + (1-\theta_j)\log\frac{1-\rho}{1-\theta_j}\right], \end{aligned} \tag{16}$$

where $\Gamma(\cdot)$ denotes the Gamma function. Please readers refer to Appendix A.4 for the detailed derivation of $\widetilde{L}(q, \boldsymbol{\xi})$. Note that the inference of the local variational vector $\boldsymbol{\xi}$ corresponds to the optimization issue $\max_{\boldsymbol{\xi}}\widetilde{L}(q, \boldsymbol{\xi})$. By computing the derivative of $\widetilde{L}(q, \boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$, we can obtain the optimal solution for each $\xi_i$ as

$$\xi_i = \sqrt{\mathbf{x}_i^{\mathrm{T}}\mathbf{D}\mathbf{x}_i},\ i = 1, \ldots, n. \tag{17}$$

The following Algorithm 1 lists the main steps to learn the parameters involved in our model. The algorithm terminates if the maximum number of iterations is reached or the change of ELBO in two successive iterations is less than a small value (say, for example, $10^{-4}$).

**Algorithm 1** Inference of model parameters : vbvs_logit(**X**, **y**)

**Input: X**, **y**

1: Initialize $a_0 = 10^{-2}$, $b_0 = 10^{-4}$, $t = 1$, $\boldsymbol{\theta}^{(0)} = \mathbf{1}$, $\boldsymbol{\xi}^{(0)} = \mathbf{0}$, $a_j = a_0 + 1/2 (j = 1, \cdots, p)$.
2: **while** the convergence criterion does not satisfy **do**
3: Update $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}$, $a_j$, $b_j$, $\theta_j$ $(j = 1, \cdots, p)$ according to (15);
4: Update $\boldsymbol{\xi}$ with (17);
5: Compute ELBO $\widetilde{L}(q, \boldsymbol{\xi})$ defined in (16);
6: Let $t = t + 1$;
7: **end while**

**Remark 1.** When fitting a logistic regression model with some data, it is usual to let $\mathbf{x}_i := [1\, \mathbf{x}_i^{\mathrm{T}}]^{\mathrm{T}}$ and assume that the model involves an intercept term $\beta_0$. In the proposed model, we impose a Gaussian prior $\beta_0 \sim \mathcal{N}(\beta_0 \mid 0, {\alpha_0}^{-1})$ on $\beta_0$, where $\alpha_0$ is let to be governed by $\mathrm{Gam}(\alpha_0 \mid a_0, b_0)$. In the meantime, we also equip $\beta_0$ with a latent indicator variable $\gamma_0$. However, we do not infer $\gamma_0$ but simply set $\theta_0 = 1$, namely, the model is supposed to always include the intercept item. In the process to infer parameters with Algorithm 1, only the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ in Eq. (15) need to be changed to include the corresponding items.

**Remark 2.** For clarity, here we summarize how to obtain the final model. Based on the outputs of Algorithm 1, let $\hat{\boldsymbol{\Gamma}} = \mathrm{diag}(\hat{\boldsymbol{\gamma}}) = \mathrm{diag}(\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$, where $\hat{\gamma}_j = \mathbb{I}(\theta_j \geq 0.5)$ and $\mathbb{I}(\cdot)$ denotes the indicator function. Then, the variables with $\hat{\gamma}_j = 1$ can be deemed as important and unimportant otherwise. Furthermore, the final model can be gained as $p(y = 1|\mathbf{x}) = \sigma(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$ in which $\hat{\boldsymbol{\beta}} = (\mu_1\hat{\gamma}_1, \mu_2\hat{\gamma}_2, \ldots, \mu_p\hat{\gamma}_p)^{\mathrm{T}}$.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

### 2.3. Choice of hyperparameters and parameter initialization

To the best of our knowledge, the effect of the initialization of $a_0$, $b_0$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ to Algorithm 1 is not significant. By following the strategy used by Drugowitsch (2013), we initialized $a_0 = 10^{-2}$, $b_0 = 10^{-4}$ via the uninformative prior fashion as stated in Algorithm 1. Regarding $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, we recommend to setting their initial values as $\boldsymbol{\theta}^{(0)} = \mathbf{1}$ and $\boldsymbol{\xi}^{(0)} = \mathbf{0}$, respectively.

As a matter of fact, the hyperparameter $\rho$ plays the role in controlling penalty strength. Hence, it is very important to take an appropriate value for it so that Algorithm 1 can achieve the best selection performance. From the updating formula for $\theta_j$ in (15), we can see that larger $\rho$ leads to larger $\theta_j$ and thus a dense model. Considering that the tuning strategy suggested by Ormerod et al. (2017) is computationally intensive for our model, we have to explore a cheaper alternative. In what follows, two schemes to determine $\rho$ will be discussed.

#### 2.3.1. Grid search of $\rho$

A popular way to tune $\rho$ is grid search, that is, selecting its optimal value from the gridded feasible region according to a certain criterion. Under the variational Bayesian framework, there are generally three criteria helping us to select the best model. The naive criterion is ELBO, i.e., $\widetilde{L}(q, \boldsymbol{\xi})$. The other two criteria are Bayesian information criterion (BIC) and variational BIC (VBIC) (You et al., 2014) whose definitions are

$$\mathrm{BIC} = 2\sum_{i=1}^{n} \log\left[1 + \exp\left(-y_i \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\beta}}\right)\right] + \mathrm{tr}(\hat{\boldsymbol{\Gamma}}) \log n \tag{18}$$

and

$$\mathrm{VBIC} = -2\widetilde{L}(q, \boldsymbol{\xi}) + 2\mathbb{E}_q \log p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}), \tag{19}$$

respectively. In Eq. (18), $\mathrm{tr}(\hat{\boldsymbol{\Gamma}})$ means the trace of the matrix $\hat{\boldsymbol{\Gamma}} = \mathrm{diag}(\hat{\gamma}_1, \ldots, \hat{\gamma}_p)$ where $\hat{\gamma}_j = \mathbb{I}(\theta_j > 0.5)$. Particularly, we sampled 100 values equivalently spaced on the linear scale between $-10$ to 3 and let them be a candidate set for $\log \frac{\rho}{1-\rho}$. Then, the optimal $\rho$ was set to be the one minimizing a certain criterion (either BIC, VBIC or negative ELBO).

To study how well the above criteria perform to set $\rho$, we utilized two simulated toy examples to do some experiments. (i) In Example 1, there are $p = 30$ variables and each observation was drawn from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. The sample size $n$ was set to be $n = 25, 100, 200, 500$, respectively. The response vector $\mathbf{y}$ was randomly drawn from the Bernoulli distribution with success rate $1/(1 + \exp(-\mathbf{X}\boldsymbol{\beta}))$. With respect to the true coefficient values, only six covariates with indices 1, 6, 11, 16, 21, 26 were assumed to be important and their coefficients were taken as $-2$, $-1.5$, $-1$, 1, 1.5, 2, respectively. For the rest unimportant ones, the coefficients were all set as zero. (ii) In Example 2, we considered a more complicated situation in which the design matrix $\mathbf{X}$ was drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$. And there are $p = 50$ covariates, while only the 1st, 11st, ..., 41st ones, with coefficients 0.8, are important. Under this model, we considered the cases with $n = 35, 50, 100, 200$ samples. Compared with Example 1, the coefficients for the truly important covariates in Example 2 are weak. Plus the high correlation between important and unimportant covariates, it is more challenging to accurately identify important variables.

To illustrate how the estimates of coefficients vary when $\rho$ changes from 0 to 1, Fig. 2 depicts their trace plots that are obtained with several sample sizes for a simulation of Example 1. In each subplot, the dash lines indicate the coefficient estimates obtained with the optimal $\rho$'s provided by three criteria. It is obvious that ELBO always favors a dense model. As for VBIC, it selects very sparse models when $n = 25$ and 100, while dense models when $n = 200$ and 500. In contrast, the models identified by BIC are always closest to the true model regardless of sample size.

Aiming at getting more insights about the behavior of each criterion, Fig. 3 shows the inferred posterior probabilities of each indicator variable as well as the estimates of each coefficient for a simulation of Example 1 with $n = 100$. Here, four other Bayesian methods ICMM (Pungpapong et al., 2015), MCMC (Tüchler, 2008), VBIS (Carbonetto and Stephens, 2012) and EMVS (Mcdermott et al., 2016) were also included into comparison. It can be observed from Fig. 3 that ICMM, VBIS and VBVS-VBIC miss 3, 2 and 3 important variables, respectively. Although EMVS chooses the six truly important variables, it wrongly includes five unimportant variables. The performance of VBVS-ELBO is the worst since it generates too many false positives. By contrast, VBVS-BIC and MCMC only falsely exclude one important variable. Furthermore, the mean parameter biases (*MPB*, please refer to the formula (26) in Section 3 for its definition) created by VBVS-BIC, VBVS-VBIC, VBVS-ELBO, ICMM, MCMC, VBIS and EMVS are 0.0507, 0.1613, 0.0608, 0.2090, 0.3090, 0.0618 and 0.2303, respectively. Due to limited space, the corresponding results for Example 2 are not shown here since they are similar to those demonstrated in Figs. 2 and 3.

For Examples 1 and 2, we repeated the simulations 100 times and recorded some evaluation metrics (i.e., $F_1$, *Acc.* and *MPB*, the definitions are provided in Section 3) as shown in Table 1. Although ELBO's prediction accuracy and coefficient estimation improve as $n$ increases, ELBO is unable to select important variables and actually it always includes all the variables. As a result, the $F_1$ score of ELBO keeps unchanged for each sample size. Table 1 shows that BIC and VBIC can select sparse models, but VBIC performs worse than BIC with regard to almost all metrics.

Based on the previous analysis, we can come to a conclusion that BIC outperforms VBIC and ELBO in terms of variable selection. In the meantime, the *Acc.* and *MPB* of BIC are very competitive with those of VBIC and ELBO as manifested in Table 1. As for the main reasons why BIC outperforms VBIC and ELBO, it can be explained as below. Notice that BIC only contains two

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

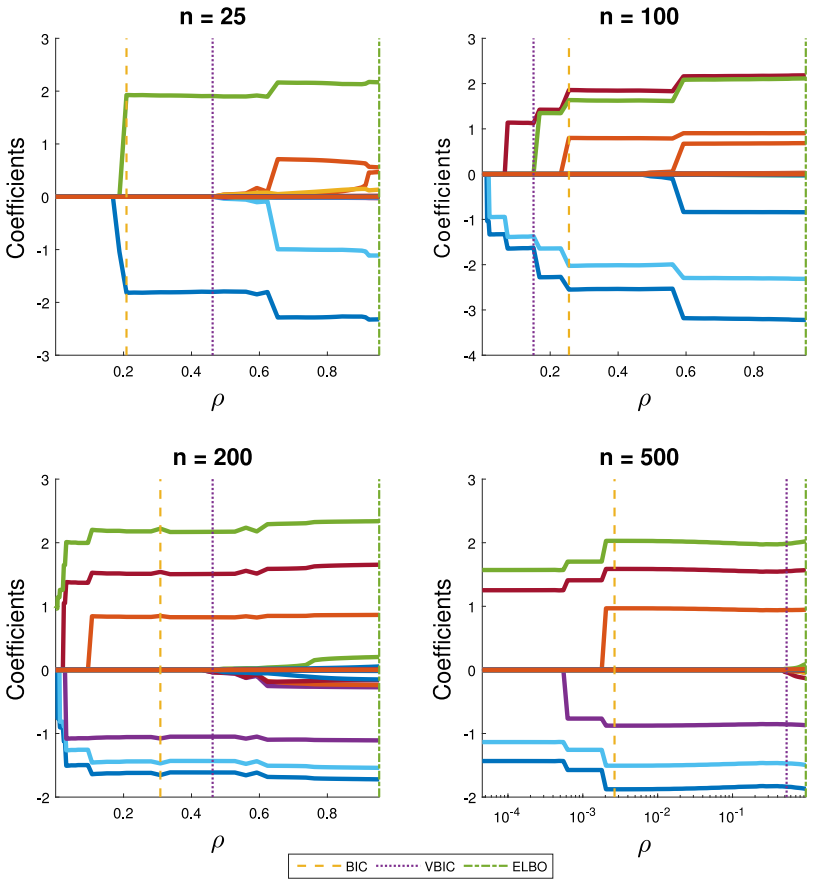



**Fig. 2.** Coefficient trace plots for Example 1 with different sample sizes. The dashed lines in each subplot indicate the models with the optimal $\rho$'s being tuned by BIC, VBIC and ELBO, respectively.

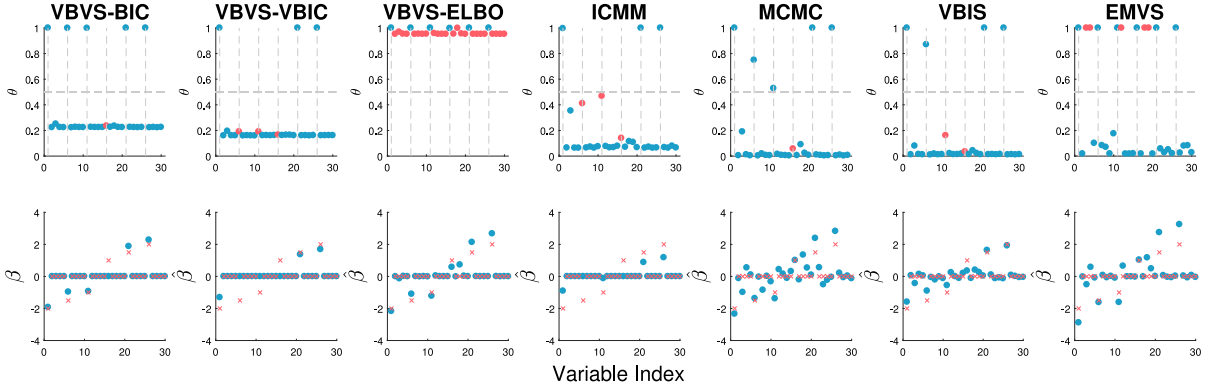



**Fig. 3.** The first row: The posterior probabilities of $\hat{\gamma}_j = 1\,(j = 1, \ldots, p)$ (that is, $\theta_j$) produced by each algorithm. The six vertical dashed lines in each subplot indicate truly important variables and the horizontal dashed lines correspond to the threshold value 0.50. The blue and red points stand for the variables which are, respectively, identified correctly and wrongly. The second row: the coefficients estimated by each algorithm. In each subplot, the blue circles and red crosses indicate the estimated and true values of coefficients, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

terms, that is, the measures of goodness of model fitting and model sparsity, which are exactly what we are interested in. While VBIC and ELBO include not only these two terms, but also the information of other items such as $\boldsymbol{\alpha}$ and $\boldsymbol{\xi}$. In a word, it is appropriate to tune $\rho$ in our model with BIC. In later experiments, we thus adopt BIC to select the optimal $\rho$ for our proposed VBVS method and abbreviate it as VBVS-BIC.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

**Table 1**
The performance of four criteria to set $\rho$. The criterion with larger $F_1$, larger *Acc.* and smaller *MPB* is better.

| Example 1 | $n = 25$ | | | | $n = 100$ | | | | $n = 200$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BIC | VBIC | ELBO | BBP | BIC | VBIC | ELBO | BBP | BIC | VBIC | ELBO | BBP | BIC | VBIC | ELBO | BBP |
| $F_1$ | 0.305 | 0.273 | 0.333 | 0.009 | 0.888 | 0.465 | 0.333 | 0.790 | 0.970 | 0.553 | 0.333 | 0.960 | 0.987 | 0.725 | 0.333 | 0.977 |
| *MPB* | 0.458 | 0.442 | 0.572 | 0.482 | 0.083 | 0.265 | 0.133 | 0.131 | 0.035 | 0.149 | 0.056 | 0.039 | 0.010 | 0.011 | 0.015 | 0.009 |
| *Acc.* | 0.622 | 0.611 | 0.658 | 0.504 | 0.827 | 0.699 | 0.825 | 0.795 | 0.855 | 0.770 | 0.849 | 0.848 | 0.859 | 0.858 | 0.856 | 0.858 |
| Example 2 | $n = 35$ | | | | $n = 50$ | | | | $n = 100$ | | | | $n = 200$ | | | |
| | BIC | VBIC | ELBO | BBP | BIC | VBIC | ELBO | BBP | BIC | VBIC | ELBO | BBP | BIC | VBIC | ELBO | BBP |
| $F_1$ | 0.202 | 0.178 | 0.182 | 0.020 | 0.304 | 0.257 | 0.182 | 0.133 | 0.523 | 0.462 | 0.182 | 0.604 | 0.749 | 0.477 | 0.182 | 0.945 |
| *MPB* | 0.131 | 0.095 | 0.261 | 0.066 | 0.124 | 0.100 | 0.179 | 0.061 | 0.074 | 0.063 | 0.088 | 0.039 | 0.035 | 0.032 | 0.042 | 0.008 |
| *Acc.* | 0.625 | 0.609 | 0.655 | 0.507 | 0.654 | 0.644 | 0.677 | 0.542 | 0.704 | 0.708 | 0.713 | 0.669 | 0.743 | 0.745 | 0.741 | 0.756 |

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

*2.3.2. Imposing a prior on $\rho$*

From the Bayesian viewpoint, a natural procedure to determine $\rho$ is to impose a prior on it. In some Bayesian variable selection literature (Scott and Berger, 2010; Rossell and Telesca, 2017; Rossell and Rubio, 2018), researchers often employ a beta-binomial prior assumption, that is, $\boldsymbol{\gamma}|\rho \sim \text{Binomial}(p, \rho)$ and $\rho \sim \text{Beta}(a, b)$, to avoid selecting $\rho$ manually or by grid search. Borrowing the idea into our proposed VBVS method, $\rho$ is assumed to be a latent variable that is governed by the Beta distribution $\rho \sim \text{Beta}(\rho \mid c_0, d_0)$, where $c_0, d_0$ are hyperparameters. In this manner, $\rho$ can be automatically inferred by the variational method.

For VBVS-BBP model, let the variational distribution be $q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \rho) = q(\boldsymbol{\beta})q(\boldsymbol{\gamma})q(\boldsymbol{\alpha})q(\rho)$. According to Eq. (8), the variational distribution $q(\rho)$ can be obtained as (please refer to Appendix B for the details)

$$q(\rho) = \text{Beta}(\rho \mid c, d), \tag{20}$$

where

$$c = c_0 + \sum_{j=1}^{n} \theta_j, \quad d = d_0 + \sum_{j=1}^{n} (1 - \theta_j). \tag{21}$$

Correspondingly, it can be derived that $q(\gamma_j) = \text{Ber}(\gamma_j \mid \theta_j)$ (see Appendix B.2 for the details), where

$$\theta_j = \sigma\left(\frac{1}{2}\mu_j \mathbf{y}^{\mathrm{T}}\mathbf{X}_j - s_{jj}d_{jj} - 2\sum_{i\neq j} s_{ij}d_{ij}\theta_i + \psi(c) - \psi(d)\right), \tag{22}$$

where $\psi(\cdot)$ stands for the digamma function. As for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, their variational distributions kept unchanged according to the derivation process listed in Appendix A. Therefore, the inference of the new model VBVS-BBP can be done by repeatedly updating the variational distributions of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\rho$.

Notice that the VBVS-BBP model can automatically determine $\rho$, but it involves two hyperparameters. In the uninformative fashion, the choice of $c_0 = 1$ and $d_0 = 1$ corresponds to the uniform prior. As a result, VBVS-BBP will be less efficient to yield sparse solution. Generally speaking, small $c_0$ and large $d_0$ favor a sparse model. In high-dimensional situations, we thus recommend to taking $c_0 = 1$ and $d_0 = p$ since Castillo and Vaart (2012) have shown that this choice can yield optimal posterior concentration rates in sparse settings. To verify the efficiency of this scheme to adaptively tune $\rho$, we applied VBVS-BBP to Examples 1 and 2 in Section 2.3.1 and the obtained results were reported in Table 1. It can be observed that in terms of all metrics, the performance of VBVS-BBP improves as the sample size $n$ increases. In comparison with VBVS-BIC, VBVS-BBP performs better if training data are sufficient, while it is exceeded by VBVS-BIC otherwise.

*2.4. Predictive distribution*

Besides detecting important variables, it is also desirable to make good predictions with the inferred model. To achieve this purpose, we substitute $p(\boldsymbol{\beta}|\mathcal{D})$ with the finally obtained variational distribution $q(\boldsymbol{\beta})$ and apply the lower bound to sigmoid function. In this way, we have

$$\begin{aligned}
\log p(y_i = 1|\mathbf{x}_i, \mathcal{D}) &= \log \iint p(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathcal{D}) \mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\gamma} \\
&\approx \log \iint p(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\gamma} \\
&\geq \log \iint \sigma(\xi_i) \exp\left(\frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{x}_i - \xi_i}{2} - \delta(\xi_i)\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}x_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\beta} + \delta(\xi_i)\xi_i^2\right) q(\boldsymbol{\beta}, \boldsymbol{\gamma})\, \mathrm{d}\boldsymbol{\beta}\mathrm{d}\boldsymbol{\gamma} \\
&= \log \sigma(\xi_i) + \delta(\xi_i)\xi_i^2 - \frac{\xi_i}{2} + \frac{1}{2}\left[\log \frac{|\boldsymbol{\Sigma}_N|}{|\boldsymbol{\Sigma}|} + \left(\boldsymbol{\mu}_N^{\mathrm{T}}{\boldsymbol{\Sigma}_N}^{-1}\boldsymbol{\mu}_N - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right],
\end{aligned} \tag{23}$$

where

$$\begin{aligned}
{\boldsymbol{\Sigma}_N}^{-1} &= \boldsymbol{\Sigma}^{-1} + 2\delta(\xi_i)\boldsymbol{\Theta}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Theta}, \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\Theta}\mathbf{x}_i.
\end{aligned} \tag{24}$$

Generally speaking, we select $\xi_i$ to maximize the lower bound. In doing so, there is $\xi_i = \sqrt{\mathbf{x}_i^{\mathrm{T}}\left(\boldsymbol{\Sigma}_N + \boldsymbol{\mu}_N\boldsymbol{\mu}_N^{\mathrm{T}}\right)\mathbf{x}_i}$. Hence, the prediction can be made with (23) by iteratively updating $\boldsymbol{\Sigma}_N$, $\boldsymbol{\mu}_N$, $\xi_i$ until the change of the lower bound is negligible.

## 3. Simulation studies

This section devotes to carrying out some simulated experiments to assess the effectiveness of the novel model. In the experiments, we took into account two versions of our proposed VBVS method, that is, VBVS-BIC and VBVS-BBP.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

For comparison, sparse logistic regressions with SCAD and MCP penalties (Breheny and Huang, 2011) were included. As for SCAD and MCP, they were implemented with the R package `ncvreg` and the binomial deviance estimated by the 10-fold cross validation was used to tune their penalty parameter. We also considered four Bayesian methods including ICMM (Pungpapong et al., 2017) and EMVS (Mcdermott et al., 2016), MCMC (Tüchler, 2008) and VBIS (Carbonetto and Stephens, 2012). In VBVS-BIC, 100 values which are equivalently spaced on the interval $[-10, 3]$ were set as the candidate values for $\log \frac{\rho}{1-\rho}$. Then, $\rho$ was set to be the value that minimizes BIC. With regard to VBVS-BBP, a beta distribution Beta$(1, p)$ was taken as the prior for $\rho$ and the inference method described in Section 2.3.2 and Appendix B was employed to determine $\rho$. Regarding the compared approaches, VBIS was implemented with the Matlab package `varbvs`.[1] The algorithms ICMM, EMVS and MCMC were implemented with the R packages `icmm`, `BinaryEMVS` and `BoomSpikeSlab`, respectively. By following the suggestion of Pungpapong et al. (2017), ICMM took the result of lasso as its initial value. And the penalty parameter involved in lasso was determined by the 10-fold cross-validation. The MCMC algorithm was run for 10000 iterations with an additional 1000 burn-in iterations. The parameters of the other algorithms were set as the default values. All the experiments were conducted with R or Matlab R2017a and ran on a computer with Intel Core CPU 3.60 GHz, 8.00 GB RAM and Windows 10 (64-bit) system.

For each algorithm, we focused on its selection behavior and prediction performance. For the former, we utilized the $F_1$ score defined below to achieve the purpose, that is,

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{25}$$

where TP, FP and FN are the number of true positives, false positives and false negatives, respectively. To assess one method's prediction performance, the classification accuracy (abbreviated as *Acc.* subsequently) was adopted. Generally speaking, both $F_1$ and *Acc.*, range from 0 to 1 and higher values are preferred. In addition, we also consider the mean parameter bias (abbreviated as *MPB*) of the estimated regression coefficients, namely,

$$MPB = \frac{1}{p} \sum_{j=1}^{p} \left( \beta_j - \mathbb{I}(\hat{\gamma}_j = 1) \hat{\beta}_j \right)^2, \tag{26}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. In the following experiments, we learned the parameters in each model with a training set and reported the obtained $F_1$ and *MPB* for selection assessment. Aiming at evaluating the prediction accuracy of the inferred model, a test set of size 10000 was generated independently. Regarding each test instance $\mathbf{x}$, the formula (23) was used to estimate $p(y = 1|\mathbf{x})$. With a threshold value 0.5, the label of $\mathbf{x}$ can be predicted as 1 if $p(y = 1|\mathbf{x}) \geq 0.5$ and $-1$ otherwise. On the test set, we then estimated the average accuracy to evaluate the prediction behavior of a method.

In this paper, we primarily considered the following four cases. In the simulations, all synthetic data were generated from the logit model, i.e., $\text{logit}(u_i) = \log\left(\frac{u_i}{1-u_i}\right) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, and the corresponding response $y_i$ was sampled from $\text{Ber}(y_i|u_i)$.

- **Scenario 1**: In this example, we considered a situation in which the signal strength is strong. The true coefficient vector was taken as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ where $\beta_j = 2.5$ if $j = 1, 36, 71$ and 0 otherwise. Obviously, only 3 variables are important and the rest ones are unimportant. The covariates were generated from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ii} = 1$ and $\boldsymbol{\Sigma}_{ij} = 0.94^{|i-j|} (\forall i \neq j)$. Hence, the correlation between important and unimportant covariates is strong, while the correlation between important covariates is weak. The dimensionality $p$ was set as $p = 100$ and the training sample size was set as $n = 50, 80, 110$, respectively.
- **Scenario 2**: To examine how well each algorithm will work in less sparser situations, we increased the number of important variables to 7 and let $\beta_j = 2.5$ if $j = 1, 16, 31, 46, \ldots, 91$ and 0 otherwise. Furthermore, the covariates were generated from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ii} = 1$ and $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$ $(\forall i \neq j)$. The other settings are the same as those in scenario 1.
- **Scenario 3**: Here, we considered a situation in which there exists weak signal strength. In particular, we set $\beta_j = 0.6$ if $j = 1, 16, 31, 46, \ldots, 91$ and 0 otherwise. The other settings are the same as those in scenario 2.
- **Scenario 4**: Finally, a high-dimensional model was used to compare all the considered methods. Here, we took $p = 300$ and $n = 200, 300, 400$, respectively. The predictors were created from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{ii} = 1$ and $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$ $(\forall i \neq j)$. For the true coefficient vector, only 15 elements are non-zero. Specifically, $\beta_j$ for the variables $1, 21, 41, \ldots, 181$ was set as 0.6 while $\beta_j$ for the variables 201, 221, 241, 261, 281 were taken as 2.

In simulations, each case was repeated 100 times by randomly generating 100 training sets. Although the compared algorithms were implemented with different softwares, all the algorithms were made to be trained and tested on the same data sets so that the comparison is fair. Then, the mean of $F_1$, *MPB* and *Acc.* values were reported in Table 2. In order to compare the efficiency of all algorithms, the last column of Table 2 lists their running time (measured in seconds) averaged over 100 simulations. From the experimental results, the following conclusions can be yielded.

[1] http://pcarbo.github.io/varbvs.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

**Table 2**
The performance of each method on synthetic data sets. The best results are highlighted in bold typeface, while the second best results are typed in italic typeface.

| Scenario 1 | $n = 50$ | | | | $n = 80$ | | | | $n = 110$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time |
| SCAD | 0.3724 | *80.88%* | 0.1865 | 0.4693 | 0.5218 | *84.63%* | *0.1193* | 0.6910 | 0.5610 | 86.31% | 0.1097 | 0.9337 |
| MCP | 0.3951 | 80.19% | *0.1813* | *0.4149* | 0.5122 | 84.04% | 0.1516 | *0.5930* | 0.5379 | 85.64% | 0.1356 | 0.8473 |
| ICMM | 0.2662 | 68.06% | **0.1733** | 0.1972 | **0.6409** | 83.73% | **0.1108** | 0.2837 | **0.6768** | 86.21% | *0.1027* | *0.8362* |
| MCMC | 0.3071 | 73.16% | 0.3623 | 4.0940 | 0.5806 | 79.85% | 0.3326 | 4.8501 | *0.6758* | 83.46% | 0.1928 | 5.0520 |
| EMVS | **0.4597** | 76.13% | 0.2267 | 2.8244 | 0.5823 | 83.44% | 0.2106 | 5.1129 | 0.5167 | 84.60% | 0.2156 | 6.9532 |
| VBIS | 0.2297 | 64.92% | 0.1900 | 1.8296 | 0.6323 | 84.05% | 0.1198 | 0.8555 | 0.6749 | **86.97%** | **0.0990** | 0.6968 |
| VBVS-BIC | *0.4150* | **81.04%** | 0.3074 | 1.1299 | *0.6336* | **85.12%** | 0.2013 | 1.2460 | 0.6744 | *86.75%* | 0.1401 | 1.6480 |
| VBVS-BBP | 0.2780 | 69.31% | 0.1866 | **0.0598** | 0.5457 | 80.23% | 0.1720 | **0.0631** | 0.6508 | 85.96% | 0.1488 | **0.0651** |
| Scenario 2 | $n = 50$ | | | | $n = 80$ | | | | $n = 110$ | | | |
| | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time |
| SCAD | **0.4054** | **71.83%** | **0.3805** | 0.5049 | 0.5827 | *79.69%* | 0.3030 | 0.8041 | 0.6875 | 84.02% | 0.2188 | 0.9511 |
| MCP | 0.3556 | 68.66% | *0.3875* | 0.4236 | 0.5901 | 78.73% | 0.3089 | *0.6366* | 0.7174 | 83.33% | 0.2249 | *0.7593* |
| ICMM | 0.0803 | 54.29% | 0.4254 | 0.1514 | 0.3524 | 66.99% | 0.3757 | 0.2195 | *0.7464* | 80.31% | 0.2485 | 0.4193 |
| MCMC | 0.2825 | 67.51% | 0.6491 | 4.7402 | 0.4812 | 75.99% | 0.4895 | 6.5060 | 0.6522 | 82.50% | 0.2922 | 8.2621 |
| EMVS | 0.3173 | 67.24% | 0.4187 | 2.5955 | **0.6440** | **81.77%** | **0.2780** | 4.6781 | 0.7300 | *85.33%* | **0.1792** | 6.6656 |
| VBIS | 0.0839 | 54.93% | 0.4303 | 1.8886 | 0.3668 | 67.46% | 0.3659 | 1.8877 | 0.6976 | 81.38% | 0.2194 | 0.9308 |
| VBVS-BIC | *0.3567* | *70.95%* | 0.4417 | 1.1301 | *0.6264* | *80.89%* | *0.2850* | 1.2642 | **0.7715** | **85.35%** | *0.1922* | 1.6402 |
| VBVS-BBP | 0.2139 | 64.61% | 0.4255 | **0.0593** | 0.3226 | 67.20% | 0.3794 | **0.0653** | 0.5985 | 75.89% | 0.2800 | **0.0660** |
| Scenario 3 | $n = 50$ | | | | $n = 80$ | | | | $n = 110$ | | | |
| | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time |
| SCAD | *0.1739* | 58.32% | 0.0266 | 0.6063 | **0.2709** | 62.06% | **0.0252** | 1.0711 | *0.3833* | *65.30%* | **0.0228** | 1.6341 |
| MCP | 0.1460 | 56.96% | 0.0295 | *0.5024* | 0.2366 | 60.55% | 0.0268 | *0.9223* | 0.3674 | 63.08% | *0.0243* | *1.4497* |
| ICMM | 0.0142 | 50.85% | **0.0259** | 0.2125 | 0.0532 | 52.59% | *0.0257* | 0.3525 | 0.1319 | 55.45% | 0.0253 | 0.9741 |
| MCMC | 0.0687 | 54.64% | 0.0291 | 4.6468 | 0.0475 | 52.38% | 0.2701 | 4.6303 | 0.1072 | 54.62% | 0.0283 | 5.3675 |
| EMVS | 0.1347 | 56.46% | 0.1567 | 3.8522 | 0.2563 | *62.08%* | 0.3455 | 7.0932 | 0.3064 | 64.17% | 0.4330 | 10.1868 |
| VBIS | 0.0125 | 51.12% | 0.0269 | 3.4052 | 0.0612 | 52.75% | 0.0259 | 3.2772 | 0.1596 | 56.38% | 0.0259 | 1.9500 |
| VBVS-BIC | **0.1944** | **60.85%** | 0.0957 | 1.9410 | *0.2632* | **63.33%** | 0.0623 | 2.1585 | **0.4070** | **65.83%** | 0.0478 | 2.8210 |
| VBVS-BBP | 0.1250 | *58.45%* | *0.0264* | **0.0579** | 0.1512 | 60.24% | 0.0277 | **0.0604** | 0.2944 | 61.68% | 0.0272 | **0.0648** |
| Scenario 4 | $n = 200$ | | | | $n = 300$ | | | | $n = 400$ | | | |
| | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time |
| SCAD | 0.4725 | **82.89%** | **0.0256** | 2.9580 | 0.5370 | **84.82%** | **0.0166** | 5.3024 | 0.5529 | **85.96%** | **0.0132** | 8.1160 |
| MCP | *0.4954* | *81.84%* | 0.0318 | *2.2113* | *0.5740* | *84.30%* | *0.0189* | 4.3378 | *0.6126* | *85.77%* | *0.0137* | 6.5450 |
| ICMM | 0.4768 | 81.46% | 0.0326 | 0.7942 | 0.5313 | 82.85% | 0.0264 | 1.1627 | 0.5633 | 83.99% | 0.0227 | 3.4821 |
| MCMC | 0.4747 | 80.74% | 0.0919 | 22.4098 | 0.5283 | 82.70% | 0.0822 | 43.7602 | 0.5657 | 84.10% | 0.0156 | 49.9121 |
| EMVS | 0.4867 | 80.69% | 0.0501 | 15.9963 | 0.4993 | 82.56% | 0.0526 | 25.6802 | 0.4873 | 83.65% | 0.0515 | 36.1403 |
| VBIS | 0.4533 | 81.05% | 0.0316 | 2.6640 | 0.5072 | 82.29% | 0.0264 | *2.9514* | 0.5536 | 83.64% | 0.0214 | *4.1291* |
| VBVS-BIC | **0.5146** | 81.64% | 0.0651 | 33.2313 | **0.5774** | 84.09% | 0.0409 | 39.3097 | **0.6198** | 85.62% | 0.0282 | 51.5076 |
| VBVS-BBP | 0.4820 | 81.84% | *0.0244* | **0.3877** | 0.5515 | 83.15% | 0.0190 | **0.4118** | 0.5958 | 84.63% | 0.0188 | **0.4765** |

(i) The results of scenario 1 manifest that VBVS-BIC outperforms all the other methods with regard to $F_1$ score and *Acc.* on average. When $n = 50$, EMVS and VBVS-BIC achieve the best and second best selection performance, but the prediction accuracy of VBVS-BIC is significantly higher than that of EMVS. In addition, ICMM performs worse in terms of both $F_1$ score and *Acc.* As the training size $n$ becomes larger, ICMM and VBVS-BIC achieve comparable $F_1$ scores, while the prediction accuracy of VBVS-BIC is larger. Even though the variable selection and prediction performance of MCMC can be ranked in the middle, it consumes more time than other algorithms but EMVS. Moreover, the penalty-based methods do not show much superiority over VBVS because they falsely include some unimportant variables.

(ii) When the true model includes more important variables, VBVS-BIC still works better than all other methods on average, as illustrated by the results of scenario 2. There is no one method which can perform best over all sample sizes. Howbeit, VBVS-BIC is always either the best or the second best performer.

(iii) When the signal strength is weak in scenario 3, ICMM, MCMC and VBIS exhibit poor performance in terms of variable selection and prediction, compared with SCAD, MCP, EMVS and two versions of VBVS. In this situation, VBVS-BIC is observed to always make better predictions than the other methods. Meanwhile, its ability to detect important variables is the best or very competitive with others. Thus, VBVS-BIC has stronger ability to deal with weak signal scenarios.

(iv) Note that the setting of scenario 4 is very complicated. Under this circumstance, there is no one method which consistently behaves best for all studied sample sizes in terms of all metrics. The results for this scenario show that VBVS-BIC is good at variable selection whilst SCAD works best in prediction. At the same time, MCP strikes a balance between variable selection and prediction. One little pity is that in this case, VBVS-BIC consumes more time than other algorithms because it needs to compute the inverse of a covariance matrix in each iteration.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

**Table 3**
The performance of each method on scenario 1, but with random permutation of the order of coefficients.

| Scenario 1 | $n = 50$ | | | | $n = 80$ | | | | $n = 110$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time | $F_1$ | *Acc.* | *MPB* | Time |
| SCAD | 0.3679 | *81.00%* | **0.1591** | 0.4901 | 0.5267 | *84.61%* | *0.1141* | 0.7034 | 0.5722 | *86.34%* | *0.1089* | *0.9300* |
| MCP | 0.3678 | 79.73% | 0.1822 | 0.4080 | 0.5107 | 83.97% | 0.1657 | 0.5830 | 0.5187 | 85.53% | 0.1466 | *0.8296* |
| ICMM | 0.3000 | 70.12% | *0.1677* | *0.2098* | **0.6494** | 83.71% | **0.1083** | *0.2854* | 0.6674 | 86.12% | **0.1044** | 0.9501 |
| MCMC | 0.2964 | 73.61% | 0.3737 | 4.1140 | 0.5981 | 80.73% | 0.3526 | 4.7327 | **0.6977** | 84.72% | 0.3613 | 5.0853 |
| EMVS | *0.4097* | 76.13% | 0.2267 | 2.8784 | 0.5821 | 83.46% | 0.2106 | 5.0855 | 0.5166 | 84.60% | 0.2169 | 7.2476 |
| VBIS | 0.2023 | 64.15% | 0.1933 | 1.0316 | 0.5547 | 83.43% | 0.1647 | 0.9365 | 0.5644 | 85.76% | 0.1434 | 1.4653 |
| VBVS-BIC | **0.4150** | **81.04%** | 0.3074 | 1.2414 | *0.6336* | **85.12%** | 0.2013 | 1.4607 | *0.6744* | **86.75%** | 0.1401 | 1.5846 |
| VBVS-BBP | 0.2780 | 69.31% | 0.1866 | **0.0598** | 0.5457 | 80.23% | 0.1720 | **0.0631** | 0.6508 | 85.96% | 0.1488 | **0.0651** |

(iv) In general, VBVS-BIC works better than VBVS-BBP in terms of variable selection and prediction accuracy. If the sample size is sufficiently large, VBVS-BBP outperforms SCAD and EMVS. Furthermore, it is obvious that VBVS-BBP possesses a significant advantage over all the other methods, that is, fast calculation speed.

In a word, VBVS-BIC provides the highest or second highest $F_1$ scores in all considered cases. When comes to predict new data, VBVS-BIC achieves the best prediction 6 times and the second best prediction 3 times. Meanwhile, its computational burden is satisfactory as demonstrated in Table 2. Therefore, it can be considered as a very competitive tool to resolve both variable selection and prediction tasks in high-dimensional logistic regression models.

Sometimes, the performance of a variable selection approach may be greatly affected by the order of coefficients. To verify whether this phenomenon occurs for our proposed model as well as other methods, we conducted some experiments by generating data in scenario 1 in the following manner. In each simulation, we first generated the design matrix $\mathbf{X}$ and randomly permuted the order of coefficients. Then, the columns in $\mathbf{X}$ were adjusted in line with the new order of coefficients. After doing this, the correlation between covariates as well as the information of truly important variables remains the same. By using the same other experimental setting as before, Table 3 reports the results for each algorithm. Through comparing the results with those listed in Table 2, it can be found that two versions of VBVS are very robust to the permutation since their results almost remain the same under each sample size. Howbeit, the performance of VBIS is influenced to a certain degree. As for the other algorithms, they are affected somewhat when $n = 50$ while becoming robust for larger sample sizes. The robustness of VBVS to permutation mainly comes from the fact that it simultaneously infers all $\beta_j$'s when estimating $\boldsymbol{\beta}$ while the other algorithms employ the cyclic coordinate optimization strategy to estimate $\boldsymbol{\beta}$.

## 4. Publicly available data experiments

In this section, we employed several publicly available data sets to investigate the performance of the studied methods. Here, four binary classification data sets were considered and the main characteristics of these data sets are summarized in the last two columns of Table 4. Note that the Heart and LSVT sets are available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/, while the Colon and Leukemia sets can be downloaded from http://www.ntu.edu.sg/home/elhchen/data.htm. Originally, there are 2000 and 7129 variables in Colon and Leukemia, respectively. By setting the significance level as 0.05, we applied the Wilcoxon rank-sum test to Colon so that the variables having not significant role to differ normal persons from patients are filtered out. With the significance level 0.005, the same procedure was applied to Leukemia data. As a result, the Colon and Leukemia sets have 386 and 985 variables in our experiments, respectively. Since the Heart set only has 13 variables, we included the interaction and quadratic terms in the experiments (totally 103 input variables) to increase its dimensionality.

Analogously to previous simulations, *Acc.* was utilized to assess the prediction performance of a method. Since there is no means to know which variables are truly important in real applications, we utilized *sparsity* (i.e., the fraction of number of zero coefficients to $p$) here to measure the interpretability produced by each method. In order to eliminate the randomness of data splitting to the relative performance of each method, the 10-fold cross validation was repeated 10 times to compute *Acc.* and sparsity. With purpose to evaluate the overall performance of a method, we utilized the harmonic mean of *Acc.* and sparsity (that is, $2/(1/Acc.+1/\text{sparsity})$) and Table 4 lists the results. Furthermore, Fig. 4 depicts the boxplots of the estimated *Acc.*s and sparsities for each algorithm. According to the results obtained in last section, here we only compared VBVS-BIC (abbreviated as VBVS in this section) with other algorithms.

From Fig. 4, we can find that SCAD, MCP and EMVS are prone to choose denser models, while VBIS and ICMM tend to provide sparser ones. There is no one method which consistently makes the best prediction on all sets. On the whole, frequentist methods (i.e., SCAD and MCP) have slightly higher accuracies than Bayesian ones since they are prediction-oriented techniques. Among Bayesian models, VBVS achieves the best predictions on all the data sets. In terms of the overall performance results demonstrated in Table 4, one can observe that VBVS obviously beats its rivals and MCP takes the second place. In summary, VBVS is very competitive with other popular methods and it usually strikes a better balance between interpretability and prediction.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

**Table 4**
The performance of each method (2nd–8th columns) on publicly available data sets as well as their basic information (the last two columns). The best results are highlighted in bold typeface, while the second best results are in italic typeface.

| Data set | SCAD | MCP | ICMM | MCMC | EMVS | VBIS | VBVS | $n$ | $p$ |
|---|---|---|---|---|---|---|---|---|---|
| Heart | 0.8538 | *0.8712* | 0.8615 | 0.7987 | 0.7876 | 0.8435 | **0.8769** | 143 | 103 |
| LSVT | **0.9180** | *0.9123* | 0.9050 | 0.8194 | 0.8179 | 0.8673 | 0.9099 | 126 | 310 |
| Colon | 0.8628 | 0.8730 | 0.8046 | 0.8357 | 0.8578 | **0.9229** | *0.8828* | 62 | 386 |
| Leukemia | 0.9607 | *0.9613* | 0.9343 | 0.8954 | 0.8499 | 0.9504 | **0.9861** | 72 | 985 |
| Mean | 0.8988 | *0.9044* | 0.8763 | 0.8373 | 0.8283 | 0.8960 | **0.9139** | | |

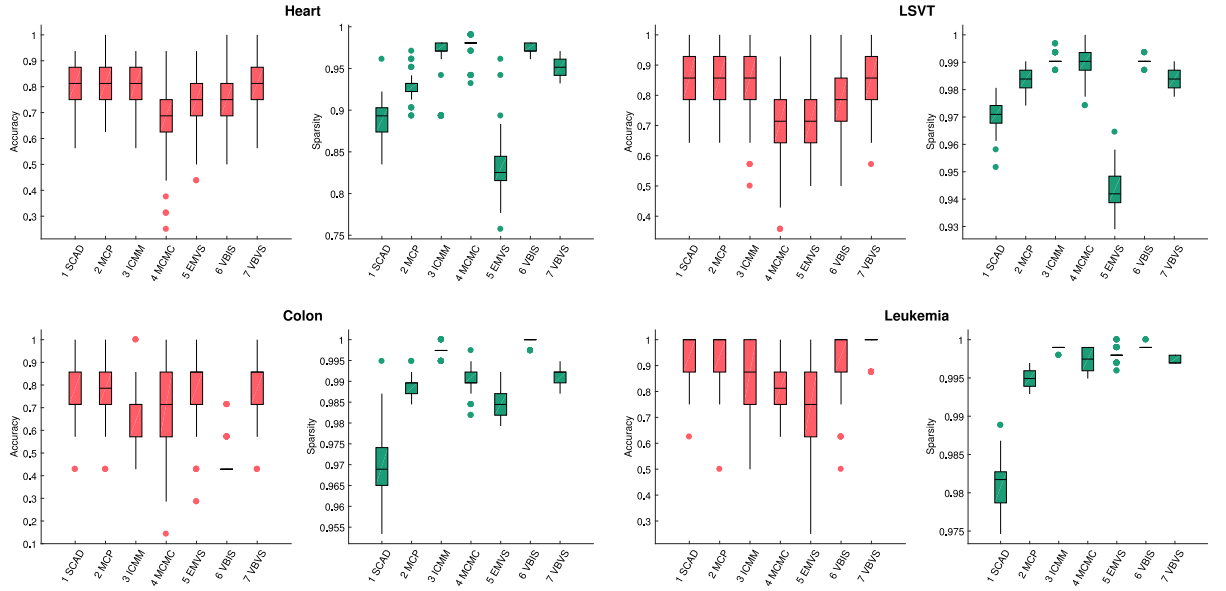


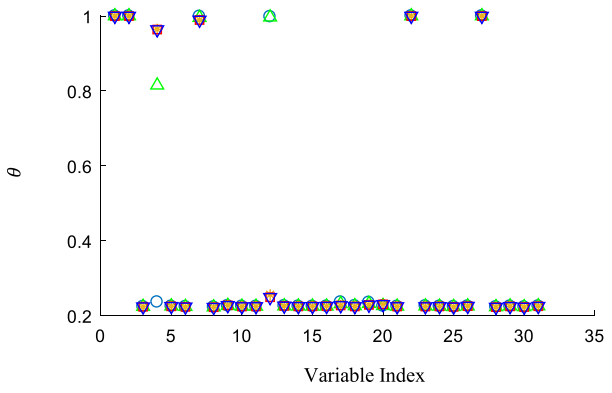

**Fig. 4.** On each studied publicly available data set, the boxplots of *Acc.* and sparsity for each algorithm.

## 5. Comparison with related models

Since the novel variational Bayesian method for logistic regression is inspired by the one developed by Ormerod et al. (2017), it is deserved to clarify the difference between two models. Besides the different model frameworks, Ormerod et al. (2017) place a prior with the same variance on $\beta_j$, i.e., $p(\beta_j) \sim \mathcal{N}(0, \sigma_\beta^2)$. And they use a deterministic annealing variant of EM algorithm and ELBO to tune the hyperparameter $\rho$ which controls sparsity. But one problem of this manner is that it tends to prefer the null model as $\sigma_\beta^2$ becomes large. On the other hand, the computational cost will be very large if the annealing EM algorithm is used to simultaneously tune $\sigma_\beta^2$ and $\rho$. As a trade-off, Ormerod et al. (2017) directly set $\sigma_\beta^2 = 10$. Nevertheless, this setting limits its application. When the coefficients of truly important variables are small, Ormerod et al. (2017)'s method is very likely to miss them. In contrast, we place priors with distinct precisions to each coefficient $\beta_j$, i.e., $p(\beta_j) \sim \mathcal{N}(0, \alpha_j{}^{-1})$ and place a Gamma hyper-prior on $\alpha_j$, i.e., $p(\alpha_j) \sim \text{Gam}(a_0, b_0)$. As a result, the variance of $\beta_j$ can be automatically determined by data and VBVS can thus avoid the preference of the null model. Furthermore, the experiments in Section 3 show that VBVS works well whenever regression coefficients are small or large. In the second aspect, it is reported that Ormerod et al. (2017)'s method is sensitive to the initialization of $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}^{(0)}$. Particularly, $\boldsymbol{\theta}$ will remain small if $\boldsymbol{\theta}^{(0)}$ is small. However, VBVS is very robust to the value of $\boldsymbol{\theta}^{(0)}$. To illustrate this finding, here we took Example 1 in Section 2.3.1 as an instance. Fig. 5 reports the convergence values of $\boldsymbol{\theta}$ with five different initializations, namely, $\boldsymbol{\theta}^{(0)} = s\mathbf{1}$, where $\mathbf{1}$ is a $p$-dimensional vector with element 1 and $s$ takes value 1, 0.75, 0.5, 0.25, 0, respectively. It is shown that VBVS is able to correctly detect important variables even though each entry in $\boldsymbol{\theta}^{(0)}$ is very small and there is no significant difference among different initializations.

Another closely related work is the indicator model proposed by Carbonetto and Stephens (2012), i.e., the VBIS algorithm mentioned in previous discussions. In VBIS, the authors adopt the spike-and-slab prior for $\boldsymbol{\beta}$ and they do not use a fully variational inference. Instead, the importance sampling strategy is employed to estimate $\boldsymbol{\theta}$. The simulated experiments in Section 3 have demonstrated that our VBVS method outperforms VBIS in terms of variable selection and prediction performance, but VBIS possesses a slight advantage in some cases when evaluating them with *MPB* value and running speed.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.




| $\theta^{(0)}$ | ○ 1 | △ 0.75 | * 0.5 | □ 0.25 | ▽ 0 |
|---|---|---|---|---|---|
| VBIC | -138.57 | -136.02 | -139.79 | -139.69 | -139.96 |
| BIC | 82.18 | 82.40 | 85.34 | 85.35 | 85.35 |
| ELBO | -148.83 | -149.17 | -149.86 | -149.86 | -149.86 |

**Fig. 5.** The convergence values of $\boldsymbol{\theta}$ with different initializations $\boldsymbol{\theta}^{(0)}$ in Example 1 of Section 2.3. The horizontal axis indicates the variable index and the vertical axis denotes the convergence values of $\boldsymbol{\theta}$ which represents the posterior probabilities of each variable being important. Meanwhile, we also list the final VBIC, BIC and ELBO values for different initializations.

## 6. Conclusions

Bayesian methods have exhibited good performance in coping with variable selection tasks in various models. Due to the advantage to easily incorporating many kinds of prior knowledge into established models, it has become an active research area in statistical modeling with the help of some efficient approximate inference tools such as variational Bayes. On account of intractable computation of MCMC in high dimension, we propose in this paper a novel variational Bayesian method for identifying important variables in high-dimensional logistic regression models. In the framework of indicator models, each covariate is equipped with a binary latent variable. Regarding the latent indicator variable, we put a Bernoulli-type prior on it. As for the specification of the hyperparameter in the Bernoulli prior, we provide two schemes to determine its optimal value so that the novel model can achieve sparsity adaptively. To infer from the posterior distributions of parameters and make predictions, one efficient variational Bayesian method is developed to make inference. The experiments on both synthetic and some publicly available data sets demonstrate the effectiveness of the novel method. In comparison with several other popular techniques, our proposed approach generally strikes a better balance between interpretability and prediction. Furthermore, its computational cost is acceptable under high-dimensional situations. As a result, it can be deemed as a competitive tool to address the variable selection problems in real high-dimensional logistic regression models.

## Acknowledgments

The authors would like to thank the editor, the associate editor and anonymous reviewers for their useful suggestions which greatly helped to improve the paper. The research of Zhang C. X. is supported by the National Natural Science Foundation of China [grant number 11671317]. The research of Zhang J. S. is supported by the National Key Research Development Program of China [grant number 2018YFC0809001], and the National Natural Science Foundation of China [grant number 61572393].

## Appendix A. Details of variational inference

In this section, we provide more details about the variational inference of the (hyper)parameters which are used in their updating formula (15) and the derivation of ELBO for our model $\widetilde{L}(q, \boldsymbol{\xi})$ in (16).

### A.1. Inference of $\boldsymbol{\beta}$

In the light of the formulae (8), (10) and (12), we can have

$$\begin{aligned}\log q(\boldsymbol{\beta}) &= \mathbb{E}_{-\boldsymbol{\beta}}\{\log h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) + \log p(\boldsymbol{\beta}|\boldsymbol{\alpha})\} + \text{constant} \\ &= \mathbb{E}_{-\boldsymbol{\beta}}\left\{\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\beta}\right\} + \text{constant} \\ &= \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbb{E}_q[\boldsymbol{\Gamma}]\mathbf{X}^{\mathrm{T}}\mathbf{y} - \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\left(\mathbb{E}_q[\mathbf{A}] + 2\mathbb{E}_q[\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}]\right)\boldsymbol{\beta} + \text{constant},\end{aligned} \tag{A.1}$$

where $\mathbb{E}_q[\cdot]$ means taking expectation with respect to the variational distribution. It is worthwhile that the constant term in (A.1) embodies all items which do not involve $\boldsymbol{\beta}$. From (A.1), it can be seen that the variational distribution of $\boldsymbol{\beta}$ is Gaussian. In what follows, we denote it by $\mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} &= \mathbb{E}_q[\mathbf{A}] + 2\mathbb{E}_q\left[\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}\right], \\ \boldsymbol{\mu} &= \frac{1}{2}\boldsymbol{\Sigma}\mathbb{E}_q[\boldsymbol{\Gamma}]\mathbf{X}^{\mathrm{T}}\mathbf{y}. \end{aligned} \tag{A.2}$$

Next, we will prove that $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\mu}$ in (A.2) have the expression as shown in the updating formula (15). The core is to compute the expectation of some items with regard to the variational distribution.

- Note that $\mathbf{A}$ is diagonal, there is thus $\mathbb{E}_q[\mathbf{A}] = \mathrm{diag}(\mathbb{E}_q[\boldsymbol{\alpha}])$. In Appendix A.2, we will see that the variational distribution of $\alpha_j$ is $\mathrm{Gam}(a_j, b_j)$. Therefore, we have $\mathbb{E}_q[\mathbf{A}] = \mathrm{diag}\left(\frac{a_1}{b_1}, \ldots, \frac{a_p}{b_p}\right)$.
- The computation of $\mathbb{E}_q[\boldsymbol{\Gamma}]$ can be done similarly. On account of $\gamma_j \sim \mathrm{Ber}(\gamma_j \mid \theta_j)$, we have $\mathbb{E}_q[\boldsymbol{\Gamma}] = \boldsymbol{\Theta}$, where $\boldsymbol{\Theta} = \mathrm{diag}(\boldsymbol{\theta})$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$.
- Since $\boldsymbol{\Gamma}$ is a diagonal matrix, there is $\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma} = \mathbf{S} \odot (\boldsymbol{\gamma}\boldsymbol{\gamma}^T)$. Moreover, it is easy to obtain that $\mathbb{E}_q[\gamma_i\gamma_j] = \theta_i\theta_j$ $(i \neq j)$ and $\mathbb{E}_q[\gamma_j^2] = \theta_j = \theta_j^2 + \theta_j(1-\theta_j)$. We can thus obtain $\mathbb{E}_q[\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}] = \mathbf{S} \odot \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \boldsymbol{\theta}\boldsymbol{\theta}^T + \boldsymbol{\Theta} \odot (\mathbf{I}_p - \boldsymbol{\Theta})$.

In summary, there is

$$\begin{aligned} \boldsymbol{\Sigma} &= \left[\mathrm{diag}\left(\frac{a_1}{b_1}, \ldots, \frac{a_p}{b_p}\right) + 2\mathbf{S} \odot \boldsymbol{\Omega}\right]^{-1}, \\ \boldsymbol{\mu} &= \frac{1}{2}\boldsymbol{\Sigma}\boldsymbol{\Theta}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \end{aligned} \tag{A.3}$$

### A.2. Inference of $\alpha_j$

Similar to the derivation of $\log q(\boldsymbol{\beta})$, for the hyperparameter $\alpha_j$ there is

$$\begin{aligned} \log q(\alpha_j) &= \mathbb{E}_{-\alpha_j}\left\{\log p(\beta_j|\alpha_j) + \log p(\alpha_j)\right\} + \text{constant} \\ &= \mathbb{E}_{-\alpha_j}\left\{\frac{1}{2}\log\alpha_j - \frac{\beta_j^2}{2}\alpha_j + (a_0 - 1)\log\alpha_j - b_0\alpha_j\right\} + \text{constant} \\ &= \left(a_0 - \frac{1}{2}\right)\log\alpha_j - \left(b_0 + \frac{1}{2}\mathbb{E}_q[\beta_j^2]\right)\alpha_j + \text{constant}. \end{aligned} \tag{A.4}$$

The above (A.4) implies that the variational distribution of $\alpha_j$ is a Gamma distribution, say, $\mathrm{Gam}(\alpha_j \mid a_j, b_j)$ with its shape and scale parameters $a_j$ and $b_j$ as

$$\begin{aligned} a_j &= a_0 + \frac{1}{2}, \\ b_j &= b_0 + \frac{1}{2}\mathbb{E}_q[\beta_j^2]. \end{aligned} \tag{A.5}$$

As discussed in Appendix A.1, there is $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$. As a result, we can have $(\boldsymbol{\beta} - \boldsymbol{\mu})(\boldsymbol{\beta} - \boldsymbol{\mu})^T \sim \mathrm{Wishart}(1, \boldsymbol{\Sigma})$, where 1 denotes the degree of freedom and $\boldsymbol{\Sigma}$ is the location parameter of the Wishart distribution. Then, there is

$$\mathbb{E}_q[(\boldsymbol{\beta} - \boldsymbol{\mu})(\boldsymbol{\beta} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma} \Rightarrow \mathbb{E}_q[\boldsymbol{\beta}\boldsymbol{\beta}^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T \equiv \mathbf{D}. \tag{A.6}$$

In addition, it is evident that $\mathbb{E}_q[\beta_j^2] = d_{jj} = \Sigma_{jj} + \mu_j^2$, $\mathbb{E}_q[\beta_i\beta_j] = d_{ij} = \Sigma_{ij} + \mu_i\mu_j$. Thus,

$$b_j = b_0 + \frac{1}{2}d_{jj}. \tag{A.7}$$

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

### A.3. Inference of $\gamma_j$

Since $\gamma_j$ is assumed to obey a Bernoulli distribution, it is easy to see that $\gamma_j^2$ and $\gamma_j$ have the same distribution. In short, $p(\gamma_j^2 = i) = p(\gamma_j = i), i = 0,1$. Hence, we have

$$\begin{aligned}
\log q(\gamma_j) &= \mathbb{E}_{-\gamma_j}\{\log h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) + \log p(\boldsymbol{\gamma})\} + \text{constant} \\
&= \mathbb{E}_{-\gamma_j}\left\{\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}\boldsymbol{\beta} + \gamma_j \log\frac{\rho}{1-\rho}\right\} + \text{constant} \\
&= \mathbb{E}_{-\gamma_j}\left\{\frac{1}{2}\beta_j\gamma_j\mathbf{X}_j^{\mathrm{T}}\mathbf{y} - s_{jj}\beta_j^2\gamma_j^2 - 2\sum_{i\neq j}s_{ij}\beta_i\beta_j\gamma_i\gamma_j + \gamma_j\log\frac{\rho}{1-\rho}\right\} + \text{constant} \\
&= \gamma_j\mathbb{E}_{-\gamma_j}\left\{\frac{1}{2}\beta_j\mathbf{X}_j^{\mathrm{T}}\mathbf{y} - s_{ij}\beta_j^2 - 2\sum_{i\neq j}s_{jj}\beta_i\beta_j\gamma_i + \log\frac{\rho}{1-\rho}\right\} + \text{constant} \\
&= \gamma_j\left\{\frac{1}{2}\mu_j\mathbf{y}^{\mathrm{T}}\mathbf{X}_j - s_{ij}d_{jj} - 2\sum_{i\neq j}s_{ij}d_{ij}\theta_i + \log\frac{\rho}{1-\rho}\right\} + \text{constant}.
\end{aligned} \tag{A.8}$$

Therefore, the variation distribution of $\gamma_j$ is still a Bernoulli distribution. In what follows, we denote it by $\mathrm{Ber}(\gamma_j \mid \theta_j)$, where

$$\theta_j = \sigma(u_j),\ u_j = \frac{1}{2}\mu_j\mathbf{y}^{\mathrm{T}}\mathbf{X}_j - s_{jj}d_{jj} - 2\sum_{i\neq j}s_{ij}d_{ij}\theta_i + \log\frac{\rho}{1-\rho}. \tag{A.9}$$

### A.4. Derivation of the evidence lower bound

Now we show how to attain ELBO for our model, namely, Eq. (16). At first, we can rewrite the ELBO as

$$\begin{aligned}
\widetilde{L}(q, \boldsymbol{\xi}) &= \mathbb{E}_q\left[\log\frac{h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})p(\boldsymbol{\beta}|\boldsymbol{\alpha})p(\boldsymbol{\gamma})p(\boldsymbol{\alpha})}{q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha})}\right] \\
&= \mathbb{E}_q\left[\log h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) + \log p(\boldsymbol{\beta}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\gamma}) + \log p(\boldsymbol{\alpha})\right] \\
&\quad - \mathbb{E}_q\left[\log q(\boldsymbol{\beta}) + \log q(\boldsymbol{\gamma}) + \log q(\boldsymbol{\alpha})\right],
\end{aligned} \tag{A.10}$$

on the basis of the assumption $q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = q(\boldsymbol{\beta})q(\boldsymbol{\gamma})q(\boldsymbol{\alpha})$. In what follows, we will compute the items in (A.10) one by one. For the first item, according to (12) we have

$$\mathbb{E}_q\left[\log h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi})\right] = \mathbb{E}_q\left\{\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}\boldsymbol{\beta} + \sum_{i=1}^{n}\left[\log\sigma(\xi_i) - \frac{\xi_i}{2} + \delta(\xi_i)\xi_i^2\right]\right\}. \tag{A.11}$$

From Eq. (A.3), we know that

$$\mathbb{E}_q\left\{\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right\} = \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Theta}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \tag{A.12}$$

$$\begin{aligned}
\mathbb{E}_q\left\{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{S}\boldsymbol{\Gamma}\boldsymbol{\beta}\right\} &= \mathrm{tr}\left\{(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})(\mathbf{S}\odot\boldsymbol{\Omega})\right\} \\
&= \frac{1}{2}\left\{p + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \mathrm{tr}\left[(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathrm{diag}\left(\frac{a_1}{b_1}, \ldots, \frac{a_p}{b_p}\right)\right]\right\}.
\end{aligned} \tag{A.13}$$

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

Plugging (A.12) and (A.13) into (A.11), the first item in (A.10) can be rewritten as

$$\mathbb{E}_q\left[\log h(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\xi})\right] = \frac{1}{2}\left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - p + \mathrm{tr}\left[(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathrm{diag}\left(\frac{a_1}{b_1},\ldots,\frac{a_p}{b_p}\right)\right]\right\} + \sum_{i=1}^{n}\left[\log\sigma(\xi_i) - \frac{\xi_i}{2} + \delta(\xi_i)\xi_i^2\right]. \tag{A.14}$$

For the item $\mathbb{E}_q\left[\log p(\boldsymbol{\beta}|\boldsymbol{\alpha})\right]$, we can have

$$\begin{aligned}\mathbb{E}_q\left[\log p(\boldsymbol{\beta}|\boldsymbol{\alpha})\right] &= \frac{1}{2}\mathbb{E}_q\left\{\left(\sum_{j=1}^{p}\log\alpha_j\right) - p\log(2\pi) - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\beta}\right\} \\ &= \frac{1}{2}\left\{\left(\sum_{j=1}^{p}\psi(a_j) - \log b_j\right) - p\log(2\pi) - \mathrm{tr}\left[(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathrm{diag}\left(\frac{a_j}{b_j}\right)\right]\right\},\end{aligned} \tag{A.15}$$

where $\psi(\cdot)$ denotes the digamma function which is defined as the logarithmic derivation of the gamma function, i.e., $\psi(x) = \partial\log\Gamma(x)/\partial x$. As for the other items in (A.10), there are

$$\begin{aligned}\mathbb{E}_q\left[\log p(\boldsymbol{\alpha})\right] &= \sum_{j=1}^{p}\mathbb{E}_q\left\{a_0\log b_0 - \log\Gamma(a_0) + (a_0-1)\log\alpha_j - b_0\alpha_j\right\} \\ &= p[a_0\log b_0 - \log\Gamma(a_0)] + \sum_{j=1}^{p}\left[(a_0-1)(\psi(a_j)-\log b_j) - b_0\frac{a_j}{b_j}\right].\end{aligned} \tag{A.16}$$

$$\begin{aligned}\mathbb{E}_q\left[\log p(\boldsymbol{\gamma})\right] &= \sum_{j=1}^{p}\mathbb{E}_q\left\{\gamma_j\log\rho + (1-\gamma_j)\log(1-\rho)\right\} \\ &= \sum_{j=1}^{p}\theta_j\log\rho + (1-\theta_j)\log(1-\rho)\end{aligned} \tag{A.17}$$

$$\begin{aligned}\mathbb{E}_q\left[\log q(\boldsymbol{\beta})\right] &= -\frac{1}{2}\mathbb{E}_q\left\{\log(\det(\boldsymbol{\Sigma})) + p\log(2\pi) + (\boldsymbol{\beta}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right\} \\ &= -\frac{1}{2}\left[\log(\det(\boldsymbol{\Sigma})) + p\log(2\pi) + p\right].\end{aligned} \tag{A.18}$$

$$\begin{aligned}\mathbb{E}_q\left[\log q(\boldsymbol{\alpha})\right] &= \sum_{j=1}^{p}a_j\log b_j - \log\Gamma(a_j) + (a_j-1)(\psi(a_j)-\log b_j) - a_j \\ &= \sum_{j=1}^{p}(a_j-1)\psi(a_j) + \log b_j - a_j - \log\Gamma(a_j).\end{aligned} \tag{A.19}$$

$$\begin{aligned}\mathbb{E}_q\left[\log q(\boldsymbol{\gamma})\right] &= \sum_{j=1}^{p}\mathbb{E}_q\left\{\gamma_j\log\theta_j + (1-\gamma_j)\log(1-\theta_j)\right\} \\ &= \sum_{j=1}^{p}\theta_j\log\theta_j + (1-\theta_j)\log(1-\theta_j).\end{aligned} \tag{A.20}$$

Based on the above formulae, it is not difficult to verify that the $\widetilde{L}(q,\boldsymbol{\xi})$ defined in (A.10) is equivalent to the ELBO given in (16).

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

## Appendix B. Variational inference of VBVS-BBP

In this section, we discuss how to infer the VBVS-BBP model. Obviously, the Markov blanket of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ keep unchanged in VBVS-BBP. Hence, the variational distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the same as those shown in Eqs. (A.1) and (A.4). In the following discussions, we focus on describing how to infer $\boldsymbol{\gamma}$ and $\rho$ with the variational method.

### B.1. Inference of $\rho$

Based on the assumptions $\boldsymbol{\gamma} \sim \prod_{j=1}^{p} \mathrm{Ber}(\gamma_j|\rho)$ and $\rho \sim \mathrm{Beta}(c_0, d_0)$, we can deduce the variation distribution of the latent variable $\rho$ as follows, namely,

$$\begin{aligned}
\log q(\rho) &= \mathbb{E}_{-\rho}\left\{\sum_{j=1}^{p}\log p(\gamma_j \mid \rho) + \log p(\rho)\right\} + \text{constant} \\
&= \mathbb{E}_{-\rho}\left\{\left[\sum_{j=1}^{p}\gamma_j\log\rho + (1-\gamma_j)\log(1-\rho)\right] + (c_0-1)\log\rho + (d_0-1)\log\rho\right\} + \text{constant} \\
&= \left(c_0 + \mathbb{E}_{-\rho}\sum_{j=1}^{p}\gamma_j - 1\right)\log\rho + \left(d_0 + \mathbb{E}_{-\rho}\sum_{j=1}^{p}(1-\gamma_j) - 1\right)\log(1-\rho) + \text{constant} \\
&= \left(c_0 + \sum_{j=1}^{p}\theta_j - 1\right)\log\rho + \left(d_0 + \sum_{j=1}^{p}(1-\theta_j) - 1\right)\log(1-\rho) + \text{constant}.
\end{aligned} \tag{B.1}$$

On the basis of the above formula (B.1), the variational distribution $q(\rho)$ is still a Beta distribution. In what follows, we denote it by $\mathrm{Beta}(\rho \mid c, d)$, where

$$c = c_0 + \sum_{j=1}^{p}\theta_j, \quad d = d_0 + \sum_{j=1}^{p}(1-\theta_j). \tag{B.2}$$

### B.2. Inference of $\boldsymbol{\gamma}$

The inference of $\boldsymbol{\gamma}$ in VBVS-BBP can be done similarly as shown in Eq. (A.8), that is,

$$\begin{aligned}
\log q(\gamma_j) &= \mathbb{E}_{-\gamma_j}\{\log h(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) + \log p(\boldsymbol{\gamma})\} + \text{constant} \\
&= \gamma_j\mathbb{E}_{-\gamma_j}\left\{\frac{1}{2}\beta_j\mathbf{X}_j^{\mathrm{T}}\mathbf{y} - s_{ij}\beta_j^2 - 2\sum_{i\neq j}s_{ij}\beta_i\beta_j\gamma_i + \log\frac{\rho}{1-\rho}\right\} + \text{constant} \\
&= \gamma_j\left\{\frac{1}{2}\mu_j\mathbf{y}^{\mathrm{T}}\mathbf{X}_j - s_{ij}d_{jj} - 2\sum_{i\neq j}s_{ij}d_{ij}\theta_i + \mathbb{E}_{-\gamma_j}\left(\log\frac{\rho}{1-\rho}\right)\right\} + \text{constant}.
\end{aligned} \tag{B.3}$$

Note that $\mathbb{E}_{-\gamma_j}\left(\log\frac{\rho}{1-\rho}\right) = \mathbb{E}_{-\gamma_j}(\log\rho - \log(1-\rho)) = \psi(c) - \psi(c+d) - (\psi(d) - \psi(c+d)) = \psi(c) - \psi(d)$, where $\psi(\cdot)$ is the digamma function. Consequently, the variation distribution of $\gamma_j$ is still a Bernoulli distribution $\mathrm{Ber}(\gamma_j \mid \theta_j)$ with $\theta_j$ as

$$\theta_j = \sigma(u_j),\ u_j = \frac{1}{2}\mu_j\mathbf{y}^{\mathrm{T}}\mathbf{X}_j - s_{jj}d_{jj} - 2\sum_{i\neq j}s_{ij}d_{ij}\theta_i + \psi(c) - \psi(d). \tag{B.4}$$

## Appendix C. Supplementary Materials

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2018.08.025.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.

## References

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer-Verlag, New York.
Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational inference: A review for statisticians. J. Amer. Statist. Assoc. 112 (518), 859–877.
Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat. 5 (1), 232–253.
Carbonetto, P., Stephens, M., 2012. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 7 (1), 73–107.
Castillo, I., Vaart, A.V.D., 2012. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. Ann. Statist. 40 (4), 2069–2101.
Drugowitsch, J., 2013. Variational Bayesian inference for linear and logistic regression. arXiv preprint. Available at https://arxiv.org/abs/1401.1022.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Ghosh, J., Li, Y., Mitra, R., 2018. On the use of Cauchy prior distributions for Bayesian logistic regression. Bayesian Anal. 13 (2), 359–383.
Hernández-Lobato, J.M., Hernández-Lobato, D., Suárez, A., 2015. Expectation propagation in linear regression models with spike-and-slab priors. Mach. Learn. 99 (3), 437–487.
Horii, S., 2017. Sparse Bayesian logistic regression with hierarchical prior and variational inference. In: Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA.
Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. John Wiley & Sons.
Jaakkola, T., Jordan, M., 2000. Bayesian parameter estimation via variational methods. Stat. Comput. 10 (1), 25–37.
Jiang, W., Zhang, C., 2014. Paths following algorithm for penalized logistic regression using SCAD and MCP. Comm. Statist. Simulation Comput. 43 (5), 1064–1077.
Koslovsky, M.D., Swartz, M.D., Leon-Novelo, L., Chan, W., Wilkinson, A.V., 2018. Using the EM algorithm for Bayesian variable selection in logistic regression models with related covariates. J. Stat. Comput. Simul. 88 (3), 575–596.
Krishnapuram, B., Carin, L., Figueiredo, M.A.T., Hartemink, A.J., 2005. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE Trans. Pattern Anal. Mach. Intell. 27 (6), 957–968.
Kuo, L., Mallick, B., 1998. Variable selection for regression models. Sankhyā: Indian J. Stat., Ser. B 60 (1), 65–81. (1960-2002).
Kyung, M., Gill, J., Ghosh, M., Casella, G., 2010. Penalized regression, standard errors, and Bayesian lassos. Bayesian Anal. 5 (2), 369–412.
Latouche, P., Mattei, P.-A., Bouveyron, C., Chiquet, J., 2016. Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression. J. Multivariate Anal. 146, 177–190.
Liang, Y., Liu, C., Luan, X.-Z., Leung, K.-S., Chan, T.-M., Xu, Z., Zhang, H., 2013. Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. BMC Bioinformatics 14 (1), 198.
Mackay, D., 1992. The evidence framework applied to classification networks. Neural Comput. 4 (5), 720–736.
Mcdermott, P., Snyder, J., Willison, R., 2016. Methods for Bayesian variable selection with binary response data using the EM algorithm. arXiv preprint. Available at https://arxiv.org/abs/1605.05429.
Nikooienejad, A., Wang, W., Johnson, V.E., 2016. Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. Bioinformatics 32 (9), 1338–1345.
Nott, D.J., Leonte, D., 2004. Sampling schemes for Bayesian variable selection in generalized linear models. J. Comput. Graph. Statist. 13 (2), 362–382.
O'Hara, R.B., Sillanpää, .M.J., 2009. A review of Bayesian variable selection methods: What, how and which. Bayesian Anal. 4 (1), 85–117.
Ormerod, J.T., You, C., Müller, S., 2017. A variational Bayes approach to variable selection. Electron. J. Stat. 11 (2), 3549–3594.
Park, T., Casella, G., 2008. The Bayesian lasso. J. Amer. Statist. Assoc. 103 (482), 681–686.
Polson, N.G., Scott, J.G., Windle, J., 2013. Bayesian inference for logistic models using Pólya-Gamma latent variables. J. Amer. Statist. Assoc. 108 (504), 1339–1349.
Pungpapong, V., Zhang, M., Zhang, D., 2015. Selecting massive variables using an iterated conditional modes/medians algorithm. Electron. J. Stat. 9 (1), 1243–1266.
Pungpapong, V., Zhang, M., Zhang, D., 2017. Variable selection for high-dimensional generalized linear models using an iterated conditional modes/medians algorithm. arXiv preprint. Available at https://arxiv.org/abs/1707.08298.
Rossell, D., Rubio, F.J., 2018. Tractable Bayesian variable selection: Beyond normality. J. Amer. Statist. Assoc. (in press). Available at https://amstat.tandfonline.com/doi/full/10.1080/01621459.2017.1371025.
Rossell, D., Telesca, D., 2017. Nonlocal priors for high-dimensional estimation. J. Amer. Statist. Assoc. 112 (517), 254–265.
Ročková, V., 2017. Particle EM for variable selection. J. Amer. Statist. Assoc. (in press). Available at https://doi.org/10.1080/01621459.2017.1360778.
Ročková, V., George, E.I., 2014. EMVS: The EM approach to Bayesian variable selection. J. Amer. Statist. Assoc. 109 (506), 828–846.
Scott, J.G., Berger, J.O., 2010. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann. Statist. 38 (5), 2587–2619.
Spiegelhalter, D.J., Lauritzen, S.L., 1990. Sequential updating of conditional probabilities on directed graphical structures. Networks 20 (5), 579–605.
Tian, G.-L., Tang, M.-L., Fang, H.-B., Tan, M., 2008. Efficient methods for estimating constrained parameters with applications to regularized (lasso) logistic regression. Comput. Statist. Data Anal. 52 (7), 3528–3542.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Tüchler, R., 2008. Bayesian variable selection for logistic models using auxiliary mixture sampling. J. Comput. Graph. Statist. 17 (1), 76–94.
Wang, C., Blei, D.M., 2013. Variational inference in nonconjugate models. J. Mach. Learn. Res. 14, 1005–1031.
Wang, J., Liang, F., Ji, Y., 2016. An ensemble EM algorithm for Bayesian variable selection. arXiv preprint. Available at https://arxiv.org/abs/1603.04360.
Xu, Z., Chang, X., Xu, F., Zhang, H., 2012. $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. IEEE Trans. Neural Netw. Learn. Syst. 23 (7), 1013–1027.
You, C., Ormerod, J.T., Müller, S., 2014. On variational Bayes estimation and variational information criteria for linear regression models. Aust. N. Z. J. Stat. 56 (1), 73–87.
Zhang, C., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38 (2), 894–942.

Please cite this article in press as: Zhang C.-X., et al., A novel variational Bayesian method for variable selection in logistic regression models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.08.025.