Research article

Application of Bayesian variable selection in logistic regression model

Kannat Na Bangchang[1,2,*]

[1] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani, 12120, Thailand

[2] Thammasat University Research Unit in Statistical Theory and Applications, Thailand

* **Correspondence:** Email: kannat@mathstat.sci.tu.ac.th.

Abstract: Typically, in high dimensional data sets, many covariates are not significantly associated with a response. Moreover, those covariates are highly correlated, leading to a multicollinearity problem. Hence, the model is sparse since the coefficient of most covariates are likely to be zero. The classical frequentist or likelihood-based variable selection via any criterion such as Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC) or a stepwise subset selection becomes infeasible when the number of variables are large. An alternative solution is a Bayesian variable selection. In this study, we used a variable selection via a Bayesian variable selection and the least absolute shrinkage and selection operator (LASSO) method in the logistic regression model. Moreover, those methods were expanded to be applied to real datasets.

1. Introduction

Bayesian variable selection (BVS) provides intuitive probabilistic interpretations and effectively explores the model space in a stochastic way to find the model with higher posterior probabilities. This approach is called stochastic search variable selection (SSVS). There are many stochastic searching schemes that have been developed such as the Gibbs variable selection, Geweke's BVS with block updates [1], and the reverse jump Markov Chain Monte Carlo (MCMC) algorithm [2]. Moreover, the application of BVS in the setting of n << p has appeared when analyzing genetic data from the early

2000s. Most methods use hierarchical Bayesian modeling to combine the empirical variance with a local background variance associated with neighboring genes [3]. BVS has been applied to data from genome wide association studies (GWAS) that contained millions of genetic variants or single nucleotide polymorphisms (SNPs) [4].

GWAS are a type of experiment that aims to detect genetic variation that may be linked to a type of disease. The main aims of GWAS are to try to determine the genetic risk factors for a disease and to make predictions about who may be at risk of developing a particular disease [5]. One of the biggest challenges of GWAS is the extremely large potential set of variants (in millions), but with a limited sample size (typically thousands) [6]. There are many areas for applications, such as estimating the heritability, calculating genetic correlations, making clinical risk predictions, and informing drug development. The details of GWAS and the relevant biological backgrounds are explained in many textbooks on genetics. The general purpose of GWAS is to identify genomic sequence differences among individuals that phenotypically differ [6]. A genotype is the genetic makeup of an individual, whereas a phenotype is a feature (of interest) that may be a result of the physical expression of genes. The most common sequence variations in the human genome are SNPs. SNPs are defined as loci with alleles that differ at a single base, with the rarer allele having a frequency of at least 1% in a random set of individuals in a population [7]. The aim of GWAS is to detect SNPs that have a statistically significant association with the trait of interest. These SNPs are called genomic risk loci.

Variable selection involves selecting a subset of the p variables of size p' such that p << n and the matrix inverse exists. It is a challenge to either construct the model from small subsets of all variables or to choose the covariates that are associated with the response due to the small number of covariates that are likely to be associated with the response in GWAS applications. Traditional variable selection methods such as forward, backward, and stepwise selections cannot be used in this situation. Due to the multicollinearity of X (when p >> n), this leads to the objective function not being unimodal. Hence, there are many different models that would equally fit the data set. The rest of paper is organized as follows. The following sections present an overview of the Bayesian variable selection in a logistic regression model. We present the model and MCMC algorithm for the Bayesian variable selection. A simulation study is reported to investigate the selection ability in both under independent data and correlated cases. This model is applied to a real data set, and concluding remarks are given in the last section.

2. Materials and methods

2.1. The Bayesian variable selection in logistic regression model

Now, we discuss the binary regression model. The standard form is discussed here, where y_i is a binary variable ($y_i \in 0,1$; i = 1,..., n) and has a Bernoulli distribution for a collection of n objects. Additionally, we have measurements on p covariates $x_i = (x_{i1},...,x_{ip})$. The parameter in the logistic model can be denoted as $g^{-1}(v_i)$, where g is a link function, v_i is the linear predictor that equals $x_i\beta$, and β is a $p \times 1$ column vector of regression coefficients.

Albert and Chib introduced a latent variable (y_i) that has a normal prior distribution, and hence

a conjugate normal posterior distribution, where $y_i = X'\beta + \varepsilon_i$, and the error term ($\varepsilon_i$) has a standard normal distribution [8]. If β is specified through a prior distribution with a probit link, then it leads to probit regression. A conjugate normal prior distribution is selected for β, p(β) = N(**b**, v), where **b** is the prior mean vector and v is the prior covariance matrix. Usually, a zero vector **b** = 0 is chosen for the prior mean and either the prior covariance matrix $v = c^2 I_p$ (independent) or the g-prior $v = c^2 (X'X)^{-1}$.

However, the logistic regression is often more widely used over the probit model when applied to biostatistical data sets; however, due to the lack of conjugacy, this is computationally more expensive. Holmes and Held introduced a latent variable z with the conjugate normal prior that led to a simpler conjugate formulation for the logistic regression model [9]. This version of the logistic regression model is discussed below:

$$y_j = 1 \text{ if } z_j > 0,$$

$$y_j = 0 \quad \text{otherwise,}$$

$$z_j = x_j \beta + \varepsilon_j,$$

$$\varepsilon_j \sim N(0, \lambda_j),$$

$$\lambda_j = (2\phi_j)^2,$$

$$\phi_j \sim KS(i.i.d.),$$

$$\beta_j \sim p(\beta).$$

The auxiliary variables ϕ_j are independent random variables from the Kolmogorov-Smirnov (KS) distribution [10]. Andrews and Mallows proved that 2AB had a logistic distribution, where A was the normal distribution and B was the Kolmogorov-Smirnov distribution [11]. In this case, ϕ_j is generated from the independent of KS; then, $(2\phi_j)^2$ is set as λ_j and λ_j is the variance in the normal distribution. It leads to a normal scale mixture distribution for ε_j in a marginal logistic distribution. Hence, this model is equivalent to a Bayesian logistic regression model [11]. The prior distribution of β is assumed to be a normal $N(b,v)$. Then, the posterior distribution of β is normal, with the mean B and the covariance matrix V as the standard for Bayesian modelling [9]:

$$\beta \,|\, z, j \sim N(B, V),$$

$$B = V(v^{-1}b + X'\lambda^{-1}z),$$

$$V = (v^{-1} + X'\lambda^{-1}X)^{-1},$$

$$\lambda^{-1} = diag(\lambda_1^{-1}, ..., \lambda_n^{-1}).$$

Holmes and Held extended the Bayesian logistic regression model to incorporate variable selection

by including a vector of covariate indicator variables $\gamma = (\gamma_1, ..., \gamma_p)$, where $\gamma_j \in \{0,1\} (j = 1, ..., p)$ corresponds to the indicator variable in the hierarchical model for variable selection [9].

The hierarchical model setup for variable selection is described below:

$$y_{\gamma j} = 1 \text{ if } z_{\gamma j} > 0,$$

$$y_{\gamma j} = 0 \quad \text{otherwise,}$$

$$z_{\gamma j} = x_{\gamma j} \beta_\gamma + \varepsilon_j,$$

$$\varepsilon_j \sim N(0, \lambda_j),$$

$$\lambda_j = (2\phi_j)^2,$$

$$\phi_j \sim KS(i.i.d.),$$

$$\beta_j \sim N(b_\gamma, v_\gamma),$$

$$\gamma \sim p(\gamma).$$

The prior distribution $p(\gamma)$ is the product of Bernoulli distributions of the variables γ_i with prior probabilities π_i. This is given by $p(\gamma) = \prod_{i=1}^{p} \pi_i^{\gamma_i} (1 - \pi_i)^{1-\gamma_i}$.

For γ, the Bernoulli prior is set with small prior probabilities since the expected number of selected SNPs on the GWAS are small. Under the simulation studies, we choose the small constant prior probabilities $\pi_i = p^* / p$ for $i = 1, ..., p$. The expected number of covariates, denoted as p^*, is set to be small (e.g., either three or five). However, in real data sets, we do not know the exact true number of SNPs. Instead of fixing the prior probabilities, we can choose a more flexible Beta-Binomial distribution for γ using the identity $p(\gamma) = \int p(\gamma | \pi) p(\pi) d\pi$, where $p(\gamma | \pi) = \prod_{i=1}^{p} \pi_i^{\gamma_i} (1 - \pi_i)^{1-\gamma_i}$, and with a hyper-prior distribution for π that is denoted by $p(\pi) = \pi^{a-1}(1-\pi)^{b-1} / B(a,b)$, where $B(a,b)$ is a Beta function.

The prior distribution of the regression coefficient β_γ is defined for the variables for which $\gamma_i = 1$, where $(b_\gamma = 0_{p_\gamma \times 1})$ and $\gamma = c^2 I_p$, c^2 is the normalizing constant, and I_p is the identity matrix of the size $p_\gamma \times p_\lambda$. The hierarchical logistic regression model gives a joint posterior distribution for $\{\beta_\gamma, \gamma, z, \lambda\}$ that can be written as $p(\beta_\gamma, \gamma, z, \lambda | X_\gamma, y) \propto p(y | z) p(z | \lambda, \beta_\gamma, X_\gamma) p(\beta_\gamma | \gamma) p(\gamma) p(\lambda)$, where $p(\lambda_i) \sim 1/4\sqrt{\lambda_i} KS(0.5\sqrt{\lambda_i})$ and $p(z | \lambda, \beta_\gamma, X_\gamma) = N(X_\gamma \beta_\gamma, \lambda)$.

2.2. MCMC algorithm for Bayesian variable selection

Since there is a high correlation between parameters in a single updating, which leads to the slow

mixing of the Markov chains, Zucknick and Richardson proposed jointly updating z and λ from this model [12].

2.2.1. Updating z and λ

The first update is drawn from $p(z,\lambda \mid \beta,\gamma,X,y) = p(z \mid \beta,\gamma,X,y)p(\lambda \mid z,\beta,\gamma,X)$.

1) The inversion method can be used to draw $p(z \mid \beta,\gamma,X,y)$, the steps of which are given below:

(a) $i=1,\ldots,n$, $p(z_i \mid \beta,\gamma,X,y) \propto Logistic(x_{\gamma i},1)I(z_i > 0)$ if $y_i = 1$ and

$p(z_i \mid \beta,\gamma,X,y) \propto Logistic(x_{\gamma i},1)I(z_i \leq 0)$ if $y_i = 0$.

(b) Calculate the Cumulative Distribution Function (CDF) $F(x)$ of the logistic distribution above. Sample $u_i \sim U[0,1]$ and solve for $F(F^{-1}(u_i)) = u_i$.

2) Rejection sampling can be applied to the sample $p(\lambda \mid z,\beta,\gamma,X)$. Those steps consist of the following:

(a) Sample $u_i \sim U[0,1]$.

(b) Sample λ_i from the candidate density $g(\lambda_i) \sim GIG(0.5,1,r_i^2) = \dfrac{r_i}{IG(1,|r_i|)}$, where IG is an

Inverse Gaussian distribution with $p(x) = \sqrt{\dfrac{|r_i|}{2\pi x^3}} \exp(-\dfrac{|r_i|(x-1)^2}{2x}), x \geq 0$, where $r_i^2 = (z_i - x_{\gamma i}\beta_\gamma)^2$.

(c) If $u_i \leq \alpha(\lambda_i)$, then accept λ_i, where $\alpha(\lambda_i) = \dfrac{l(r_i^2,\lambda_i)p(\lambda_i)}{Mg(\lambda_i)}$ with

$l(r_i^2,\lambda_i) = p(z_i \mid x_{\gamma i},\beta_\gamma,\lambda_i) = N(x_{\gamma i}\beta_\gamma,\lambda_i)$, $p(\lambda_i)$ being the IG distribution.

Otherwise, reject λ_i and go back to step (a).

Moreover, the use of an alternative series expansion of $KS(0.5\lambda_i)$ in [10] gives the following:

$$\alpha(\lambda_i) = N(x_{\gamma i}\beta_\gamma,\lambda_i)\frac{1}{4\sqrt{\lambda_i}}KS(0.5\sqrt{\lambda_i}).$$

2.2.2. Updating z and λ

(β_γ,γ) are updated jointly using the following identity:

$$p(\beta_\gamma,\gamma \mid z,\lambda,X) = p(\gamma \mid z,\lambda,X)p(\beta_\gamma \mid \gamma,z,\lambda,X).$$

1) With a starting value of $\gamma = \gamma^0$, β_γ can be directly sampled from $N(B_\gamma^*,V_\gamma^*)$, where

$$B_\gamma^* = V_\gamma^* x_{\gamma^*}' \lambda^{-1} z \quad \text{and} \quad V_\gamma^* = (x_{\gamma^*}' \lambda^{-1} x_{\gamma^*} + v_{\gamma^*}^{-1}).$$

2) Then, γ is updated using the following steps of a Metropolis-Hastings algorithm:

(a) (Add/delete step) At the t-th iteration, a single covariate is selected at random and the proposal distribution is given by $q(\gamma_j^*) = 1$ if $\gamma_j = 0$ and $q(\gamma_j^*) = 0$ if $\gamma_j = 1$.

(b) The acceptance probability for updating γ is given by the following:

$$\alpha(\gamma) = \min(1, \frac{\left|V_{\gamma^*}\right|^{1/2} \left|V_{\gamma}\right|^{1/2} \exp(0.5 B'_{\gamma^*} V_{\gamma^*}^{-1} B_{\gamma^*})(1 - \pi_i)}{\left|V_{\gamma}\right|^{1/2} \left|V_{\gamma^*}\right|^{1/2} \exp(0.5 B'_{\gamma} V_{\gamma}^{-1} B_{\gamma}) \pi_i}),$$

where π_j is the prior probability that γ_j takes the value 1.

(c) Set $\gamma^t = \gamma^*$ if $\gamma_j = 0$ with probability $\alpha(\gamma)$ and $\gamma^t = \gamma^{t-1}$ otherwise.

3. Results

3.1. Simulation study

The section will describe investigations into the performance of the Bayesian variable selection in logistic regression models, where the response is binary (categorical) and the covariate settings are discussed.

3.1.1. Simulation setup

The simulation consists of p =500, 1000, and 2000 covariates alongside n =500 observations. The Binomial probability (for covariates) is set to 0.1. We set the effect size of regressors as 1, 1.5, and 2. At each setting, we generate 10 data sets. In each case, we find that the inclusion probability is very low. Hence, we set higher values in each case. The simulation involves 3 scenarios, similar to the linear regression model. x_1, x_2, and x_3 are assumed to be associated with the response, while the others are not. For the variable selection, we use the *logisticVS* function in the *bvsflex* package, setting the number of iterations as 200,000 based on the pilot runs, thus indicating that a longer time is needed for MCMC convergence. There are several parameters in the package. The prior mean for β is set to 0. We conduct sensitivity studies by changing the values of many hyper-parameters, where g for the prior covariance matrix of β is set to 0.1, 1, 10, and 100, the prior precision for β is set to 100, and the prior mean of Beta prior distribution is set to 0.06. Under the package, the additional hyper-prior distribution for p is given by $p \sim Beta(a,b)$. In addition, there are 3 classes of prior distributions for the hyper-parameter g: Inverse-gamma, hyper-g, and none. We choose none, which implies that g is assumed to be fixed at the specified value. The *bvsflex* package on R-forge is used for the variable selection in logistic regression models that is based within R, version 4.2.2 (http://bvsflex.r-forge.r-project.org) [13].

We consider a simulation scenario where we have p = 500, 1000, and 2000 covariates alongside n = 500 observations. Each explanatory variable is generated from a Binomial distribution with parameters (2, p), where p takes the value 0.01. The effect size for the regressor is 2. The dependent variable is generated from the Binomial distribution with n=1 (for the binary values). Prior to generating the response variables, the input variables (x_i) are centered to the mean 0, as this model does not contain an intercept.

With many possible combinations of settings, a few cases are chosen and are presented below. These results are shown in terms of the inclusion probability of each covariate for the logistic regression model with g = 1, both under independent and correlated cases (Tables 1 and 2).

Table 1. The inclusion probability for each covariate for the logistic regression model when g = 1 and under the independent case.

Prior mean	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
0.2	0.56	0.31	0.53	0.11	0.12	0.13	0.15	0.10	0.14	0.13	0.12	0.12
0.4	0.49	0.47	0.50	0.23	0.13	0.14	0.14	0.13	0.19	0.16	0.15	0.13
0.6	0.40	0.42	0.46	0.27	0.19	0.17	0.16	0.24	0.24	0.16	0.19	0.21
0.8	0.39	0.40	0.45	0.29	0.19	0.24	0.18	0.24	0.25	0.24	0.17	0.19

Table 2. The inclusion probability for each covariate for the logistic regression model when g = 1 under the correlated case.

Prior mean	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
0.2	0.28	0.29	0.30	0.23	0.21	0.19	0.16	0.08	0.06	0.08	0.08	0.07
0.4	0.27	0.29	0.28	0.21	0.29	0.11	0.19	0.13	0.14	0.06	0.08	0.15
0.6	0.25	0.21	0.29	0.21	0.17	0.26	0.18	0.16	0.13	0.12	0.17	0.16
0.8	0.25	0.23	0.27	0.20	0.26	0.19	0.22	0.18	0.13	0.17	0.18	0.17

As can be seen from Tables 1 and 2, the inclusion probability is high when the effect size is high for an associated covariate (i.e., x_1, x_2, and x_3), with 500 covariates under the independent case. Moreover, the inclusion probability for other covariates is less than those for the associated covariates in all situations, since the true model contains all 3 covariates.

3.2. Application

This dataset highlights a GWAS of heart disease from Prof. Sandosh Padmanabhan's lab at Cardiovascular Sciences at Glasgow. A partial analysis of the data is covered in Padmanabhan and Joe (2017). There are 5312 observations; however, after excluding rows with missing values, 5158 observations remain. Moreover, there are only 3731 that match the genotype data by the patient ID. The covariates measured include the ID, age, sex, body mass index (BMI), smoking behavior, and the existence of previous cardiovascular diseases. There are two response variables. The first is severe stage 2 hypertension, where a person either has a systolic blood pressure (SBP) greater than or equal to 140 mmHg or a diastolic blood pressure (DBP) greater than or equal to 90 mmHg. The second outcome variable is a "hypertensive crisis", when a person has a blood pressure higher than 180/120 mmHg, thus requiring urgent medical care. There are 15,221 associated covariates of SNP genotype information for each individual.

The results from Table 3 show that some covariates are selected in both methods - BVS and LASSO; a fraction of these overlap and could potentially indicate stronger signals within the data. Figure 1 shows the traceplot, autocorrelation for sampled values, and the posterior regression coefficient when using a Bayesian variable selection. As we note, sampled values of the Bayesian framework are well mixed and a present satisfactory stability for the autocorrelation.

Table 3. The summary of SNP when using Bayesian, glmnet package via $\lambda.\min$ and $\lambda.1se$.

SNP	Bayesian	$\lambda.\min$	$\lambda.1se$
rs12905013	✓	✓	✓
rs1863464	✓	✓	✓
rs1619030	✓		
rs10519226	✓		
rs2442464	✓		
rs1055879	✓		
rs745636	✓		
rs1719336		✓	✓
rs999787	✓	✓	✓
rs1718939		✓	✓
rs8039254	✓	✓	✓
rs2165488	✓		
rs1865923	✓		
rs12902710	✓	✓	✓
rs8027171	✓	✓	✓
rs12591031		✓	✓
rs1392161	✓	✓	✓
rs1719336	✓		
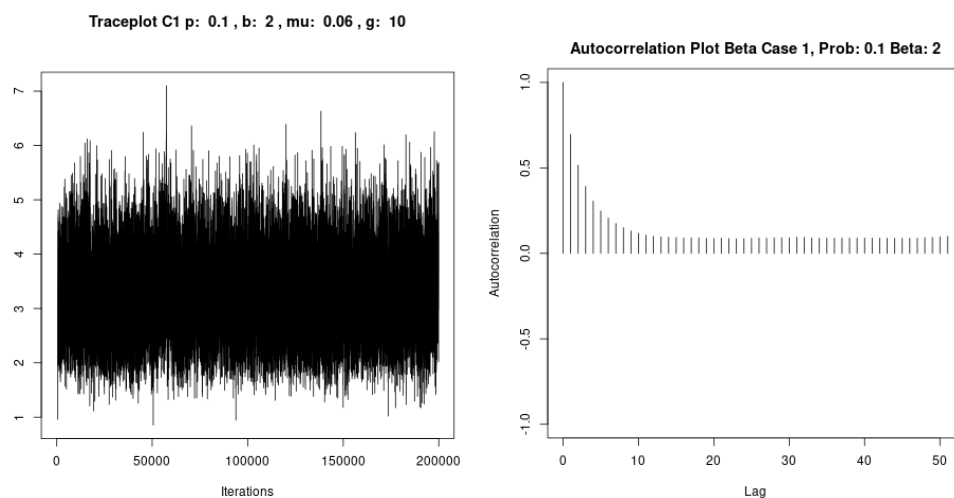
Figure 1. Traceplot and autocorrelation for regression coefficient.

4. Conclusions

Computing the marginal inclusion probability of each variable helps determine whether the variable should be included in the model. Under the case of independent covariates, the inclusion probabilities of the associated covariates appeared to increase when the effect size increased. For example, the inclusion probabilities of the associated covariates were about 0.5 when the effect size was 1, whereas the inclusion probabilities of the associated covariates were nearly 1 when the effect size was 1.5. When the effect size was low ($\beta = 0.5$), the inclusion probabilities were also low for the

associated covariates. The estimation of the regression coefficients for the associated covariates and the non-associated covariate values were accurate (similar to the true effect size). Moreover, we saw that when the Binomial proportion of the generative model (p) increased to be closer to 0.5, the inclusion probability also increased. Moreover, this study suggested higher inclusion probabilities with higher effect sizes ($\beta = 1.5, 2$). For the correlated cases, the inclusion probabilities under the low correlation were higher than that of the high correlation. For example, when $r = 0.1$, the inclusion probabilities of about 0.6 dropped to inclusion probabilities of about 0.2 when $r = 0.8$. The inclusion probabilities of the associated covariates were increased when the effect size increased. A similar result was witnessed when the probability p of the generative Binomial distribution increased. Taken together, inclusion probabilities increased as p was increased.

In the future, we will consider the Fisher information for the variable selection used in ranked set sampling from the simple linear regression model [14].

Use of AI tools declaration

The author declares he has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

Conflict of interest

The author declares no conflict of interest in this paper.

References

1. J. Geweke, Variable selection and model comparison in regression, *Bayesian Stat.*, **5** (1995). https://doi.org/10.21034/wp.539

2. P. J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82** (1995), 711–732. https://doi.org/10.2307/2337340

3. P. L. Baldi, A. D. Long, A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inference of gene changes, *Bioinformatics*, **17** (2001), 509–519. https://doi.org/10.1093/bioinformatics/17.6.509

4. J. Wakefield, Bayes factors for genome-wide association studies: Comparison with *P*-Values, *Genet. Epidemiol.*, **33** (2009), 79–86. https://doi.org/10.1002/gepi.20359

5. W. S. Bush, J. H. Moore, Genome-wide association studies, *PLoS Comput. Biol.*, **8** (2012), 1–11. https://doi.org/10.1371/journal.pcbi.1002822

6. E. Uffelmann, Q. Q. Huang, S. M. Munung, J. D. Vries, Y. Okada, A. R. Martin, et al., Genome-wide association studie, *Nat. Rev. Method. Primers*, **59** (2021), 1–21. https://doi.org/10.1038/s43586-021-00056-9

7. B. J. B. Keats, S. L. Sherman, *Population genetics, in Emery and Rimoin's principles and practice of medical genetics*, 6 Eds., London: Academic Press, 2013. https://doi.org/10.1016/b978-0-12-383834-6.00015-x

8. J. H. Albert, S. Chib, Bayesian analysis of binary and Polychotomous response data, *J. Am. Stat. Assoc.*, **88** (1993), 669–679. https://doi.org/10.1080/01621459.1993.10476321

9. C. Holmes, L. K. Held, Bayesian auxiliary variable models for binary and multinomial regression, *Bayesian Anal.*, **1** (2006), 145–168. https://doi.org/10.1214/06-ba105

10. L. Devroye, *Non-uniform random variate generation*, New York: Springer, 1986. https://doi.org/10.1007/978-1-4613-8643-8

11. D. F. Andrews, C. L. Mallows, Scale mixtures of normal distributions, *J. Roy. Stat. Soc.*, **36** (1974), 99–102. https://doi.org/10.1111/j.2517-6161.1974.tb00989.x

12. M. Zucknick, S. Richardson, MCMC algorithms for Bayesian variable selection in the logistic regression model for large-scale genomic applications, *arXiv Computation*, 2014.

13. M. Zucknick, 2013. Available from: https://r-forge.r-project.org/projects/bvsflex/.

14. S. Wang, W. Chen, R. Yang, Fisher information in ranked set sampling from the simple linear regression model, *Commun. Stat.-Simul. C.*, **53** (2024), 1274–1284.