International Journal of Data Science and Analysis
2019; 5(3): 42-51
http://www.sciencepublishinggroup.com/j/ijdsa
doi: 10.11648/j.ijdsa.20190503.12
ISSN: 2575-1883 (Print); ISSN: 2575-1891 (Online)

Data Augmentation and Bayesian Methods for Multicategory Support Vector Machines

Yeqian Liu

Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, USA

Email address:
yeqian.liu@mtsu.edu

To cite this article:
Yeqian Liu. Data Augmentation and Bayesian Methods for Multicategory Support Vector Machines. *International Journal of Data Science and Analysis.* Vol. 5, No. 3, 2019, pp. 42-51. doi: 10.11648/j.ijdsa.20190503.12

Received: June 30, 2019; **Accepted**: July 24, 2019; **Published**: August 7, 2019

Abstract: The support vector machine (SVM) has become very popular within the machine learning literature. Recently, SVM has received much attention from statisticians. It is well known that for multicategory classification problem, the commonly used multicategory SVM is based on the frequentist framework. In this paper, we develop a multi-class support vector machine under the Bayesian framework. Numerical studies were performed by EM and the Bayesian algorithm Gibbs sampler. Our results have shown that the classification accuracy of the Bayesian approach is comparable to that of frequentist approaches, while Bayesian approach also has the advantage of providing estimates of uncertainty in predictions.

Keywords: Multivariate Classifcation, MSVM, MCMC, EM

1. Introduction

Recently, Support vector machine has drawn attention from the statistics community during the high popularity of SVM rising in machine learning literature. The well-known two category SVM for the binary classification problem can be interpreted geometrically with a hyperplane which gives the maximum margin of discriminating one class from the other. See Boser, Guyon, & Vapnik [1], Vapnik [2], and Burges [3]. In two category SVM, the separation is achieved by a hyperplane which has the largest distance to the data of the two groups.

Bayesian approach, which has been developed rapidly during the past thirty years, plays a very important role in statistics. Nicholas G and Steven L [4] applied the Bayesian approach to SVM classification problem. In their paper, they developed a latent variable representation of original SVM, which enable to use EM or MCMC algorithms do parameter estimation. In their method, data augmentation methods can be formulated in terms of complete data sufficient statistics, which is a considerable advantage when working with large data sets, where most of the computational expense comes from repeatedly iterating over the data. Methods based on complete data sufficient statistics need only compute those statistics once per iteration, at which point the entire parameter vector can be updated. [4]

Recently, it was shown that the support vector machine (SVM) [5] admits a Bayesian interpretation through the technique of data augmentation. However, existing inference methods for the Bayesian support vector machine [6] can only handle two-category classification problem under Bayesian framework. Based on stochastic variational inference [7] and inducing points [8], we develop a Bayesian support vector machine for multicategory classification problem in this paper. The proposed Bayesian multicategory SVM not only inherits the advantage of robustness against outliers, advanced accuracy [9], and guaranteed error rate [10] from the frequentist formulation of the SVM, but like all Bayesian methods, it also has the advantage of modeling with high flexibility, automatic parameter tuning, and providing estimates of uncertainty in predictions.

This article is organized as follows. Section 2 states the loss function and Bayesian models for multi-SVM. Section 3 describe the Point estimation by EM and other related algorithms. Section 4, present the MCMC for SVM. Finally, Section 5 gives concluding remarks and discussion of future directions.

2. Multicategory Support Vector Machines

Recall that the standard support vector machines for the

binary case based on a k-dimensional predictors,where the class labels y_i are either 1 or -1. And the L^α-norm regularized support vector classifier seeks β minimizing:

$$\sum_{i=1}^{n} \max(1 - y_i x_i^T \beta, 0) + \nu^{-\alpha} \sum_{j=1}^{k} |\beta_j/\sigma_j|^\alpha \quad (1)$$

where σ_j is the standard deviation of the j'th predictor variable, except that $\sigma_1 = 1$ for the intercept term, and ν is a tuning parameter. [11]

2.1. Model and Notations

In this section, we will extent this model to the multicategory case. Consider a k-category classification problem with a vector of predictors $\boldsymbol{x}_i = (1, x_1, \ldots, x_{p-1})$. Class labels y_i are defined as follows. If example i falls into class j, then y_i is equal to a k-dimensional vector with 1 in the j-th coordinate and $-1/(k-1)$ elsewhere. Accordingly, define a $p-k$ function $f(x) = (f_1(x), \ldots, f_k(x))$, for any $x \in \mathbb{R}^p$. Where $f_r(x) = x^T\beta_r$ for r=1,...,k. Let L be a $k-k$ function that maps y_i to a k-dimensional vector with 0 in the j'th coordinate and 1 elsewhere, if example i is in class j. Now, we propose an extension of support vector classifier to choose a set of coefficients β to minimize:

$$d_\alpha(\beta, \nu) = \sum_{i=1}^{n} L(y_i) \times (f(x_i) - \boldsymbol{y}_i)_+ + \nu^{-\alpha} \sum_{r=1}^{k} \sum_{j=1}^{p} |\beta_{rj}/\sigma_j|^\alpha \quad (2)$$

The scaling variable σ_j is the standard deviation of the j-th predictor variable, except that σ_1=1 for the intercept term. And ν is a tuning parameter. It can be shown that minimizing (2) is equivalent to maximize the following pseudo-posterior density

$$\begin{aligned} p(\beta|\nu, \alpha, y) &\propto exp\{-d_\alpha(\beta, \gamma)\} \\ &\propto C_\alpha(\nu) L(y|\beta) p(\beta|\nu, \alpha) \end{aligned} \quad (3)$$

where $C_\alpha(\nu)$ is a pseudo-posterior normalization constant. The pseudo-likelihood $L(y|\beta)$ can be written as

$$L(y|\beta) = \prod_{i=1}^{n} L_i(y_i|\beta) = \exp\{-2 \sum_{i=1}^{n} L(y_i) \times (f(x_i) - y_i)_+\} \quad (4)$$

For simplicity, define $U_r(x_i) = (f_r(x_i) - y_{ir})$, then

$$L(y_i) \times (f(x_i) - y_i)_+ = \sum_{r=1, r\neq j}^{k} max(U_r(x_i), 0) \quad (5)$$

if example i is from class j.

Therefore, the pseudo-likelihood for example i is rewritten as

$$\begin{aligned} L_i(y_i|\beta) &= \exp\{-2 \sum_{r=1, r\neq j}^{k} \max(U_r(x_i), 0)\} \\ &= \prod_{r=1, r\neq j}^{k} exp\{-2\max(U_r(x_i), 0)\} \\ &= \prod_{r=1, r\neq j}^{k} \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{ir}}} \exp\left(-\frac{(U_r(x_i)+\lambda_{ir})^2}{2\lambda_{ir}}\right) d\lambda_{ir} \end{aligned} \quad (6)$$

The last step is from Theorem in [12] where λ_i is a vector of latent variables paired with y_i.

Denote c_i as the class label of example i, then the pseudo-likelihood can be rewritten as

$$\begin{aligned} L(y|\beta) &= \prod_{i=1}^{n} L_i(y_i|\beta) \\ &= \prod_{i=1}^{n} \prod_{r=1, r\neq c_i}^{k} \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{ir}}} \exp\left(-\frac{(U_r(x_i)^2}{2\lambda_{ir}}\right) d\lambda_{ir} \\ &= \prod_{i=1}^{n} \prod_{r=1}^{k} \int_0^\infty \left[\frac{1}{\sqrt{2\pi\lambda_{ir}}} \exp\left(-\frac{(U_r(x_i))^2}{2\lambda_{ir}}\right)\right]^{I(c_i\neq r)} \end{aligned} \quad (7)$$

Based on the objective function in (2), consider the exponential power prior distribution for β which contains the regularization penalty as follows

$$p(\beta|\nu, \alpha) = \prod_{r=1}^{k} \prod_{j=1}^{p} p(\beta_{rj}|\nu, \alpha) = \prod_{r=1}^{k} \left(\frac{\alpha}{\nu\Gamma(1+\alpha^{-1})}\right)^p \exp\left(-\sum_{j=1}^{p} \left|\frac{\beta_{rj}}{\nu\sigma_j}\right|^\alpha\right) \quad (8)$$

In general, consider $\alpha \in (0,2]$ where $\alpha = 2$ corresponds to the "ridge regrssion" and $\alpha = 1$ corresponds to the "lasso". Then the prior regularization penalty can be expressed as a scale mixture of normals.

$$p(\beta_{rj}|\nu, \alpha) = \int_0^\infty \phi(\beta_{rj}|0, \nu^2\omega_{rj}\sigma_j^2) p(\omega_{rj}|\alpha) d\omega_{rj} \quad (9)$$

where $p(\omega_{rj}|\alpha)$ is exponential with mean 2 for the special case of $\alpha = 1$.

2.2. Conditional Distribution

According to the above computations, especially equations (7) and (9), the support vector machine pseudo-posterior distribution can be expressed as the marginal of the complete data pseudo-posterior distribution as follows:

$$\begin{aligned} p(\beta, \lambda, \omega|y, \nu, \alpha) \propto & \prod_{i=1}^{n} \prod_{r=1}^{k} \left[\lambda_{ir}^{-0.5} \exp\left(-\frac{(x_i^T\beta_r - y_{ir} + \lambda_{ir})^2}{2\lambda_{ir}}\right)\right]^{I(c_i\neq r)} \\ & \times \prod_{r=1}^{k} \prod_{j=1}^{p} \omega_{rj}^{-0.5} \exp\left(-\frac{\beta_{rj}^2}{2\nu^2\omega_{rj}\sigma_j^2}\right) p(\omega_{rj}|\alpha) \end{aligned} \quad (10)$$

Define $\theta_{-r} \triangleq \{i: c_i \neq r\}$ as the set of all subjects who does not fall in class r. Rewrite the complete data pseudo-posterior distribution as

$$\begin{aligned} p(\beta, \lambda, \omega|y, \nu, \alpha) \propto & \prod_{r=1}^{k} \prod_{i\in\Theta_{-r}} \lambda_{ir}^{-0.5} \exp\left(-\frac{(x_i^T\beta_r - y_{ir} + \lambda_{ir})^2}{2\lambda_{ir}}\right) \\ & \times \prod_{r=1}^{k} \prod_{j=1}^{p} \omega_{rj}^{-0.5} \exp\left(-\frac{\beta_{rj}^2}{2\nu^2\omega_{rj}\sigma_j^2}\right) p(\omega_{rj}|\alpha) \end{aligned} \quad (11)$$

The full conditional distribution of β given λ, ω, y

According to equation (10), the full conditional distribution of β_r for any r=1,...,k is

$$p(\beta_r|\nu,\lambda_r,\omega_r,y) \propto \prod_{i\in\Theta_{-r}} \prod_{j=1}^{p} \exp\left(-\frac{(x_i^T\beta_r - y_{ir}+\lambda_{ir})^2}{2\lambda_{ir}}\right) \times \exp\left(-\frac{\beta_{rj}^2}{2\nu^2\omega_{rj}\sigma_j^2}\right) \quad (12)$$

Define the matrices $\Lambda_r = diag(\lambda_r), \Omega_r = diag(\omega_r)$, where the diagram elements of Λ_r and Ω_r are the elements of λ_r and ω_r, respectively. And $\sigma = diag(\sigma_1^2,\ldots,\sigma_p^2)$. Also let X_r denote a matrix with row i equal to, x_i^T, the predictor vector of the i'th subject in Θ_{-r}.

This model can be written in hierarchical form:

$$\begin{aligned} y_r - \lambda_r = &\quad X_r\beta_r + \Lambda_r^{\frac{1}{2}}\epsilon^{\lambda_r} \\ \beta_r = &\quad \frac{1}{\nu}\Omega_r^{\frac{1}{2}}\Sigma^{\frac{1}{2}}\epsilon^{\beta_r} \end{aligned}$$

where ϵ^{β_r} and ϵ^{λ_r} are vectors of iid standard normal deviates with dimensions matching β_r and λ_r.

Thus, for β_r has a conditional normal posterior distribution given by:

$$p(\beta_r|\nu,\lambda_r,\omega_r,y) \sim \mathcal{N}(b_r,B_r) \quad (13)$$

where =

$$B_r^{-1} = \nu^{-2}\sigma^{-1}\omega_r^{-1} + X_r^T\Lambda_r^{-1}X_r \text{ and } b_r = B_rX_r^T(y_r \times \lambda_r^{-1} - 1) \quad (14)$$

The full conditional distribution for λ_{ir} and ω_{rj} given β,ν,y Consider the conditional distribution of λ_{ir} for r=1,..,k and $i \in \Theta_{-r}$. Note that from the complete pseudo-posterior distribution we can get:

$$\begin{aligned} p(\lambda_{ir}|\beta_r,y_{ir}) &\propto \frac{1}{\sqrt{2\pi\lambda_{ir}}}\exp\left\{-\frac{1}{2}\left(\frac{(x_i^T\beta_r - y_{ir})^2}{\lambda_{ir}} + \lambda_{ir}\right)\right\} \\ &\sim \mathcal{GIG}\left\{\frac{1}{2},1,(x_i^T\beta_r - y_{ir})^2\right\} \end{aligned}$$

This implies that:

$$(\lambda_{ir}^{-1}|\beta_r,y_{ir}) \sim \mathcal{IG}(|x_i^T\beta_r - y_{ir}|^{-1},1) \quad (15)$$

For the full conditional distribution of ω_{rj}, it is proportional to the integrand in equation (9). In general this is complicated because its prior density $p(\omega_{rj}|\alpha)$ is generally not available. However, for the two special cases of $\alpha = 1$ and $\alpha = 2$ closed from solutions are available. When $\alpha = 2, p(\omega_{rj}|\beta_{rj})$ is a point mass at 1. For $\alpha = 1$, the full conditional distribution of ω_{rj} is:

$$\begin{aligned} p(\omega_{rj}|\beta_{rj},\nu) &\propto \frac{1}{\sqrt{2\pi\omega_{rj}}}\exp\left\{-\frac{1}{2}\left(\frac{\beta_{rj}^2/\nu^2\sigma_j^2}{\omega_{rj}} + \omega_{rj}\right)\right\} \\ &\sim \mathcal{GIG}\left(\frac{1}{2},1,\frac{\beta_{rj}^2}{\nu^2\sigma_j^2})\right) \end{aligned}$$

Thus similarly

$$(\omega_{rj}^{-1}|\beta_{rj},\nu) \sim \mathcal{IG}(\nu\sigma_j/|\beta_{rj}|,1) \quad (16)$$

Later we will use these distributions to develop learning algorithms.

3. Point Estimation by EM and Other Related Algorithms

In this section, we use the distributions obtained in Section 2 to construct EM-style algorithms to estimate the coefficients. First, an EM algorithm for learning β with a fixed value of the tuning parameter ν is developed. Then we develop an ECME algorithm to learn β and ν simultaneously.

3.1. Learning β with Fixed ν

With the augmented data λ and ω, the EM algorithm is a iterative method for finding posterior modes or MLEs. From equation (12), we know that the posterior distribution of the β_r's for r=1,...,k are independent. So we can estimate them separately using the EM algorithm. For β_r, the E-step and M-step are defined by

$$\begin{aligned} E-step\ Q(\beta_r|\beta_r^{(g)}) &= \int \log p(\beta_r|\nu,\lambda_r,\omega_r,y)p(\lambda_r,\omega_r|\beta_r^{(g)},\nu,y)d\lambda_r d\omega_r \\ M-step\ \beta_r^{(g+1)} &= \underset{\beta_r}{\text{argmax}} Q(\beta_r|\beta_r^{(g)}) \end{aligned} \quad (17)$$

Note that any term in $\log p(\beta_r|\nu,\lambda_r,\omega_r,y)$ that is free of β_r can be absorbed to the constant. This leaves us only the linear function of λ_{ir} and ω_{rj}. Thus, we only need to replace them with their conditional expectations $\hat{\lambda}_{ir}^{-1(g)}$ and $\hat{\omega}_{rj}^{-1(g)}$ for the calculation of function $Q(\beta_r|\beta_r^{(g)})$, given β_r and the observed data.

As we discussed before, the result for ω_{rj} would depend on the value of α. Focus on the case where $\alpha = 1$. And according to equation (16), we can obtain that:

$$\omega_{rj}^{-1(g)} = \nu\sigma_j|\beta_{rj}|^{-1}$$

Recall that the conditional posterior of β_R follows a multivariate normal distribution. Thus, the posterior mode will be the same as the posterior mean. By equations (13) and (14), we can get the following algorithm.

Algorithm: EM-SVM

Repeat the following until convergence

E-Step Given a current estimate $\beta_r = \beta_r^{(g)}$, compute:

$$\hat{\lambda}_{ir}^{-1(g)} = |x_i^T\beta_r - y_{ir}|^{-1},$$

$$\hat{\Lambda}_r^{-1(g)} = diag\left(\hat{\lambda}_r^{-1(g)}\right),$$

$$\hat{\Omega}_r^{-1(g)} = diag\left(\hat{\omega}_r^{-1(g)}\right),$$

M-Step Compute $\beta_r^{(g+1)}$ as

$$\beta_r^{(g+1)} = \left(\nu^{-2}\Sigma^{-1}\widehat{\Omega}_r^{-1(g)} + X_r^T\widehat{\Lambda}_r^{-1(g)}X_r\right)^{-1} X_r^T(y_r \times \hat{\lambda}_r^{-1(g)} - 1)$$

3.2. Learning β and ν Simultaneously

In order to learn β and γ together, the generalized expectation-conditional maximization algorithm (ECME) is used, where the last"E" represents the conditional maximization of either function. To implement the ECME algorithm, we assume a inverse gamma prior distribution for ν^{α}:

$$p(\nu^{-\alpha}) \propto (\nu^{-\alpha})^{\alpha_\nu - 1}\exp(-b_\nu\nu^{-\alpha})$$

Combine this prior with equation (8), we can find the conditional posterior density of ν given β and α

$$p(\nu^{-\alpha}|\beta,\alpha) \propto (\nu^{-\alpha})^{\frac{pk}{\alpha}+\alpha_\nu-1}\exp\left\{-\nu^{-\alpha}\left[b_\nu + \sum_{r=1}^{k}\sum_{k=1}^{p}\left|\frac{\beta_{rj}}{\sigma_j}\right|^{\alpha}\right]\right\}$$

The following algorithm can be obtained with minor modification of the EM-SVM algorithm.

Algorithm: ECME-SVM

E-Step Identical to the E-step of EM-SVM with $\nu = \nu^{(g)}$.

CM-Step Identical to the M-step of EM-SVM with $\nu = \nu^{(g)}$.

CME-Step Set

$$(\nu^{\alpha})^{(g+1)} = \frac{b_\nu + \sum_{r=1}^{k}\sum_{k=1}^{p}|\beta_{rj}^{(g)}/\sigma_j|^{\alpha}}{pk/\alpha + a_\nu - 1}$$

4. Fully Bayesian Multicategory Support Vector Machines

In the MSVM framework of [15], following [12, 13], we can find f(.) by minimizing the following penalized function when $\alpha = 1$,

$$d(\beta,\nu) = \sum_{i=1}^{n}\mathrm{L}(y_i)\cdot(\mathrm{f}(\mathrm{x}_i) - \mathrm{y}_i)_+ + \nu^{-1}\sum_{r=1}^{k}\sum_{j=1}^{p}|\frac{\beta_{rj}}{\sigma_j}| \quad (18)$$

or equivalently,

$$d(\beta,\nu) = \sum_{r=1}^{k}\sum_{i\in\ominus_{-r}}(f_r(\mathrm{x}_i) + \frac{1}{k-1})_+ + \nu^{-1}\sum_{r=1}^{k}\sum_{j=1}^{p}|\frac{\beta_{rj}}{\sigma_j}| \quad (19)$$

with constraints $\sum_{r=1}^{k} f_r(\mathbf{x}_i) = 0$ for $i = 1,\cdots,n$.

where σ_j is the standard deviation of the j'element of $\mathbf{x}$, ν is a tuning parameter and $\ominus_{-r} = \{i: c_i \neq r\}$, c_i is the classification number of observation i.

The minimization problem (19) can be viewed to find the mode of pseudo-posterior distribution from the Bayesian perspective. That is

$$p(\beta|\nu,\mathrm{y}) \propto \exp(-d(\beta,\nu)) \propto C(\nu)L(\mathrm{y}|\beta)p(\beta|\nu) \quad (20)$$

where $C(\nu)$ is a normalization constant. According to the form of the objective function, we can adopt the following likelihood function for the data and assume a exponential power prior for β as follows;

$$L(\mathrm{y}|\beta) = \prod_{i=1}^{n}L_i(y_i|\beta) = \exp\{-2\sum_{i=1}^{n}\mathrm{L}(\mathrm{y}_i)\cdot(\mathrm{f}(\mathrm{x}_i) - \mathrm{y}_i)_+\} \quad (21)$$

$$p(\beta|\nu) = \prod_{r=1}^{k}\prod_{j=1}^{p}p(\beta_{rj}|\nu) = (\prod_{r=1}^{k}\prod_{j=1}^{p}\frac{1}{2\nu\sigma_j})\exp(-\sum_{r=1}^{k}\sum_{j=1}^{p}\frac{|\beta_{rj}|}{\nu\sigma_j}) \quad (22)$$

where $[\beta_{rj}|\nu]$ follows the Laplace distribution.

Now,following [14], we assume a gamma prior on ν^{-1}, i.e.

$$p(\nu^{-1}) \propto (\nu^{-1})^{a_\nu - 1}\exp(-b_\nu\nu^{-1}) \quad (23)$$

with hyper-parameters (a_ν, b_ν). Then we use the independent Jeffreys noninformative prior, called the invariance prior, on σ_j,

$$p(\sigma_j) \propto \frac{1}{\sigma_j} \quad (24)$$

for $j = 1,\cdots,p$.

Theorem 1 Under the penalized function (19) and the priors (23) and (24), following the data augmentation approach proposed by [15], we have the following full conditional posterior distributions

$$[\beta_r|\nu,\lambda_r,\mathrm{w}_r,\mathrm{y}] \sim \mathcal{N}(\mathrm{b}_r,\mathrm{B}_r) \quad (25)$$

$$[\lambda_{ir}^{-1}|\beta_r,y_{ir}] \sim \mathcal{IG}(|\mathrm{x}_i^T\beta_r - y_{ir}|^{-1},1) \quad (26)$$

$$[w_{rj}^{-1}|\beta_{rj},\nu,\sigma_j] \sim \mathcal{IG}(\nu\sigma_j/|\beta_{rj}|,1) \quad (27)$$

$$[\nu^{-1}|\beta,\sigma_j] \sim \mathrm{Gamma}(pk + a_\nu - 1, b_\nu + \sum_{r=1}^{k}\sum_{j=1}^{p}\frac{|\beta_{rj}|}{\sigma_j}) \quad (28)$$

$$[\sigma_j|\nu,\beta] \sim \mathrm{Inv.Gamma}(k,\frac{1}{\nu}\sum_{r=1}^{k}|\beta_{rj}|) \quad (29)$$

for $i \in \ominus_{-r}; r = 1,\cdots,k$ and $j = 1,\cdots,p$. Where $\mathrm{B}_r^{-1} = \nu^{-2}\Sigma^{-1}\Omega_r^{-1} + \mathrm{X}_r^T\Lambda_r^{-1}\mathrm{X}_r$ and $\mathrm{b}_r = \mathrm{B}_r\mathrm{X}_r^T\Lambda_r^{-1}(\mathrm{y}_r - \lambda_r)$. And $\mathrm{y}_r = \{y_{ir}\}_{i\in\ominus_{-r}}, \lambda_r = \{\lambda_{ir}\}_{i\in\ominus_{-r}}, \Lambda_r = diag(\lambda_r)$, $\Omega_r = diag(\{w_{rj}\}_{j=1}^{p})$, $\Sigma = diag(\{\sigma_j^2\}_{j=1}^{p})$, 1 is the vector of 1's. X_r is a matrix with row i is $x_i, i \in \ominus_{-r}$.

Then the MCMC algorithm is developed from Theorem 1.

Algorithm: MCMC-MSVM

Draw $\beta_r^{(g+1)}$ from $\mathcal{N}(\mathrm{b}_r^{(g)},\mathrm{B}_r^{(g)})$ for $r = 1,\cdots,k$;

Draw $\lambda_{ir}^{-1(g+1)}$ from $\mathcal{IG}(|\mathrm{x}_i^T\beta_r^{(g+1)} - y_{ir}|^{-1},1)$ independently, for $r = 1,\cdots,k; i \in \ominus_{-r}$;

Draw $w_{rj}^{-1(g+1)}$ from $\mathcal{IG}(\nu^{(g)}\sigma_j^{(g)}/|\beta_{rj}^{(g+1)}|,1)$ independently, for $r = 1,\cdots,k$ and $j = 1,\cdots,p$;

Draw $\nu^{-1(g+1)}$ from $\mathrm{Gamma}(pk + a_\nu - 1, b_\nu + \sum_{r=1}^{k}\sum_{j=1}^{p}\frac{|\beta_{rj}^{(g+1)}|}{\sigma_j^{(g)}})$;

Draw $\sigma_j^{(g+1)}$ from $\text{Inv.Gamma}(k, \frac{1}{\nu^{(g+1)}}\sum_{r=1}^{k} |\beta_{rj}^{(g+1)}|)$ for $j = 1, \cdots, p$.

5. Application

5.1. Data Introduction

TIMSS represents the Trends in the International Mathematics and Science Study, which is one of the most important and largest global studies in education achievement. TIMSS was first conducted in 1995, and it collected data every four years on the achievement of fourth and eighth grade students internationally. TIMSS 2011 is the fifth in the series of assessments. 32 countries participated in the TIMSS 2011 assessments. There is enormous diversity among the TIMSS countries—in terms of economic development, geographical location, and population size. Countries participating in TIMSS aim for a sample of at least 4,500 students to ensure that there are enough respondents for each item. This study dedicates to improving teaching and learning in mathematics and science by comparing different country and exploring the environmental factors that predict students' achievement.

The TIMSS 2011 International Database includes the data from instruments that were administered to the students, their parents, their teachers, and their school principals. These include the student responses to the achievement items:

TIMSS mathematics and science and students, teacher and principals' responses to the student, home, teacher, and school background questionnaires. This is a large dataset with six files in it: school background data files, student background data files, student achievement data files, home background data files, student-teacher linkage files and teach background data files.

We used the 4^{th} grade USA data in the dataset. 12569 U.S students participated in this study. Students were administered a background questionnaire with questions related to their home background, school experiences, and attitudes toward reading, mathematics, and science. An example question is "About how many books are there in your home? (Do not count magazines, newspapers, or your school books." The student background data files contain students' responses to these questions. It includes 15 variables. Each variable may have multiple items. For example, mathematics self-concept has 7 items. We will use the mean of the items of a variable as the score. Except those questions especially designed for learning science, 13 potential predictors from the student background data files were used in the model. There are two demographic variables: sex and age; six environmental variables about recourses and parental support; five psychological variables about students; feeling and experiences. Figure 1 shows the description of the predictors and their descriptive statistics.

Variable name	Description	M	SD
ITSEX	Gender	\	\
ASDAGE	Age	10.2149667	0.4400797
ASBG04	Books at home	2.8667688	1.1727318
ASDG05S	Home study support	1.5748178	0.6025070
ASBG06A	Use computer at home	1.8165768	0.9580806
ASBG06B	Use computer at school	2.1690165	0.8554715
ASBG06C	Use computer at other places	2.8706163	1.0662095
ASBGPS	Parental support	1.6217624	0.6936827
ASBGFS	Feelings about school	1.7337753	0.7463121
ASDGSBS	Bulling experience at school	1.6784642	0.7708174
ASDGSLM	Like learning math	1.7690187	0.7752828
ASDGEML	Engage in math lessons	1.6124320	0.6104187
ASDGSCM	Confidence with math	1.7900122	0.7371308

Figure 1. *Predictors and descriptive statistics.*

Code	Description
1	Student performed below the Low International Benchmark
2	Student performed at or above the Low International Benchmark, but below the Intermediate International Benchmark
3	Student performed at or above the Intermediate International Benchmark, but below the High International Benchmark
4	Student performed at or above the High International Benchmark, but below the Advanced International Benchmark
5	Student performed at or above the Advanced International Benchmark

Figure 2. *Frequency of each group.*

Y		Frequency	Percent	Cumulative Frequency	Cumulative Percent
<400	1	500	3.98	500	3.98
400-474	2	1951	15.52	2451	19.50
475-549	3	4241	33.74	6692	53.24
550-624	4	4225	33.61	10917	86.86
>=625	5	1652	13.14	12569	100.00

Figure 3. *Percent of each group.*

Since the data we used are from the U.S which is a developed country, some of the measures might be skewed or reaching a celling effect. For example, ASBG05S ranges from 0 to 2 with a mean at 1.57. It is negatively skewed.

The predictor is student's achievement level. Students of highest achievement are in group 5 which the students of the lowest achievement are in group 1. Figure 2 and 3 shows the frequency and percent of each group.

5.2. Traditional Methods

Before applying the new method, three traditional methods were used to classify the current data.

Firstly, we used Linear Discriminant Analysis (LDA). Two types of prior were used: proportional and equal. The results showed that when proportional prior were used, the total error rate was 57.67%. The error count estimates were highest at the two ends. For group 1 and group 5, then error rate were both larger than 90%, while for group 3 and group 4, the error rates were about 40%. When equal prior were used, the total error rate was 62.16%, but the error rate were highest in the middle groups (all three groups had an error rate that was larger than 75%). The error rates at the ends were smaller, 47.80% for group1 and 34.62% for group 5.

Secondly, apply QDA on this data. As LDA, Two types of prior were used: proportional and equal. The results showed that when proportional prior were used, the resubstitution error rate was 56.23%. The error count estimates were highest at group1 (95%), while for group 3 and group 4, the error rates were about 50%. When equal prior were used, the total error rate was 58.76%, but the error rate were highest in the middle groups (all three groups had an error rate around 70%). The error rates at the ends were smaller, 43.40% for group1 and 29% for group 5. Comparing with LDA, QDA seems to perform a little better. Finally, we used a One-against-one (pairwise) SVM technique to analyze the current data. We used the ksvm function in the kernlab package. The training error was 63.96%, and the 10-fold cross-validation error rate was 74.93%. This error rate is larger than LDA and QDA. This indicates that the one-against-one (pairwise) SVM technique may not be appropriate for this dataset.

5.3. EM-algorithm

Two parts in EM algorithm were studied. We first consider fixed ν and study the relationship between different value of ν and number of non zero parameter estimates. When $\omega^{-1} = \infty$, it follows that $\beta_j = 0$, in which case we can simply ignore the jth column of covariate matrix. Figure 4 shows how the value of ν affect number of non zero parameters.

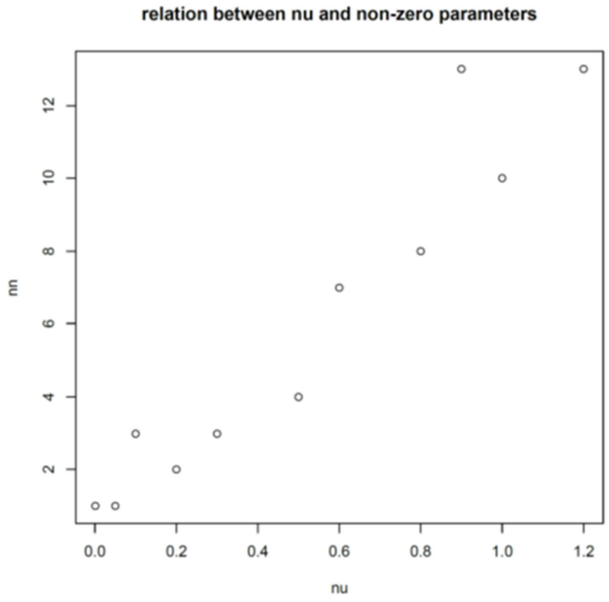

Figure 4. *Relation between ν and non zero parameters.*

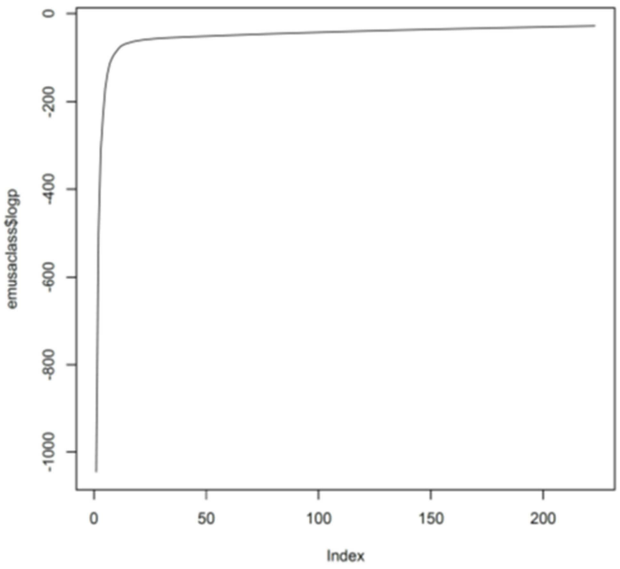

Figure 5. *Convergence of log-likelihood.*

Value of ν in Figure 1 are chosen as 0.001, 0.05, 0.1, 0.2, 0.3, 0.5, 0.6, 0.8, 0.9, 1.2. When $\nu = 0.9$ or 1.2, the number of non zero parameter are all included in classification. Vertical values are mean based on ten times computation. To study point estimate of β, we select $\nu = 0.9$. In our data set, there are 13 different covariates included, which is not too many. However, in some dataset, dimension of X could be very large. We consider a new convergence criteria, which can be called as likelihood convergence. We identity the convergence when likelihood does not change too much in iteration.

Figure 5 shows the convergence of log-likelihood. The log-likelihood increase very fast within first 20 iterations and slow after.

When estimating, β is a vector includes $(\beta_{1r}, \dots, \beta_{13r})$. We let $\lambda = 1$ and estimate of σ are computed from dataset directly. Log-likelihood convergence is one way to meet our goal, however, this criteria can not guarantee the stability of parameter estimates. Table 1 gives the results of estimation and Figure 3 shows the variance of parameter estimates. All variance are obtained based on 10 time computation. In group 1, 3, 4, β_{11} have largest variance. Variance of β_7 and β_8 in group 4 are large compared with other estimate. In group 4 and 5, variance of β_1 and β_2 are large.

Table 1. *Parameter estimation with* $\nu = 0.9,\ \lambda = 1$.

	$\boldsymbol{\beta_{i1}}$	$\boldsymbol{\beta_{i2}}$	$\boldsymbol{\beta_{i3}}$	$\boldsymbol{\beta_{i4}}$	$\boldsymbol{\beta_{i5}}$
β_{1r}	0.000	0.000	0.000	0.000	-0.160
β_{2r}	0.000	-0.503	0.560	-0.013	0.867
β_{3r}	-0.655	0.000	1.015	-0.079	0.245
β_{4r}	0.000	0.000	-0.693	-0.002	-0.643
β_{5r}	-1.037	-1.001	0.820	-0.019	0.337
β_{6r}	-0.135	-1.093	-1.310	-0.055	-0.951
β_{7r}	0.178	0.000	-0.063	0.060	0.000
β_{8r}	-2.020	0.004	-0.030	0.039	-0.103
β_{9r}	-0.896	1.118	-1.456	0.000	-0.003
β_{10r}	-0.148	0.471	-0.433	0.000	0.000
β_{11r}	-1.835	0.000	-0.002	0.032	0.000
β_{12r}	0.000	-0.393	0.000	0.000	-0.216
β_{13r}	-0.509	0.168	-0.166	0.000	0.000

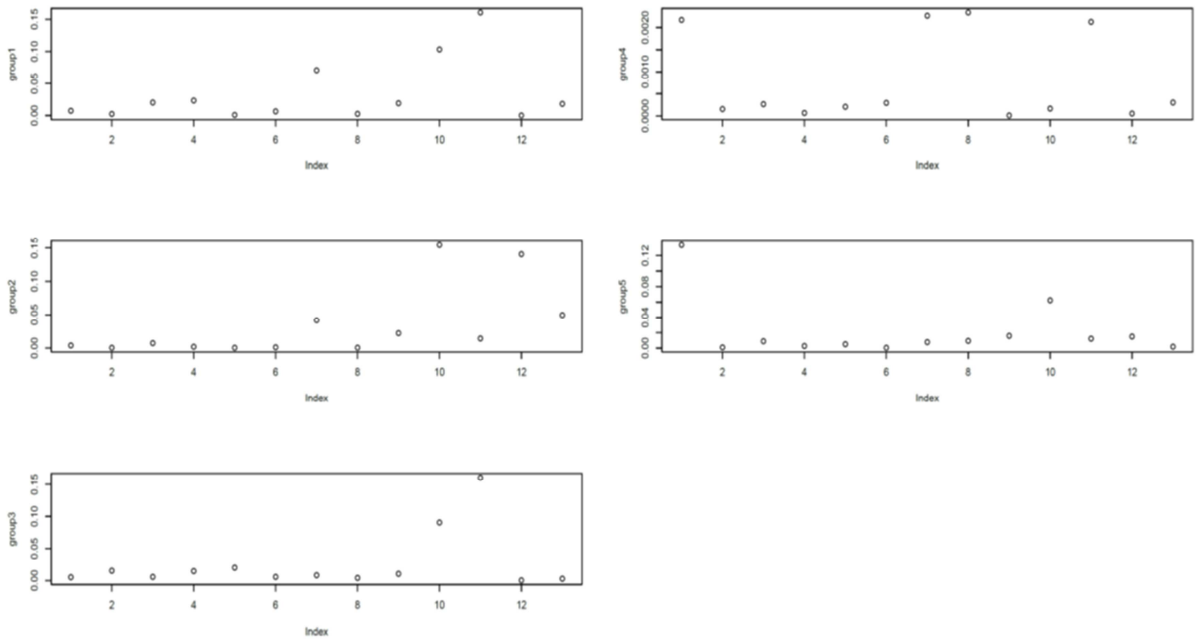

Figure 6. *Variance of Estimate.*

5.4. Bayesian MSVM

First, we test our Bayesian MSVM on a dataset , named wine, in the UCI data repository. It classifies a given silhouette as one of four types of vehicle based on 18 predictors. 500 training samples were selected and the remaining 346 samples as the testing set. All the 18 predictors were used as input variables and all the inputs are normalized to have zero mean and unit variance. We ran our MCMC algorithm for 1000 iterations and burn in 500. Figure 7 is the sample paths for the coefficients for two predictors, there is no trend among them. In fact, the sample paths of all the parameters are stationary, which indicates our MCMC is convergence. The classification error on the training set is $39\% = 195/500$, and the prediction error for the testing set is $36.99\% = 128/346$.

Then, apply our Bayesian model to the TIMSS data set, we randomly select a subset with sample size of 1000. In our procedure, 13 predictors were used as input variables and all the inputs are normalized to have zero mean and unit variance. We ran our MCMC algorithm for 1000 iterations and burn in 500. Figure 8 is the sample paths for the coefficients for two predictors, there is no trend among them. In fact, the sample paths of all the parameters are stationary, which indicates our MCMC is convergence. The classification error for the training set is $22\% = 220/1000$, and the prediction error for the testing set is $26.6\% = 133/500$. It indicates that the variable selection works pretty well.

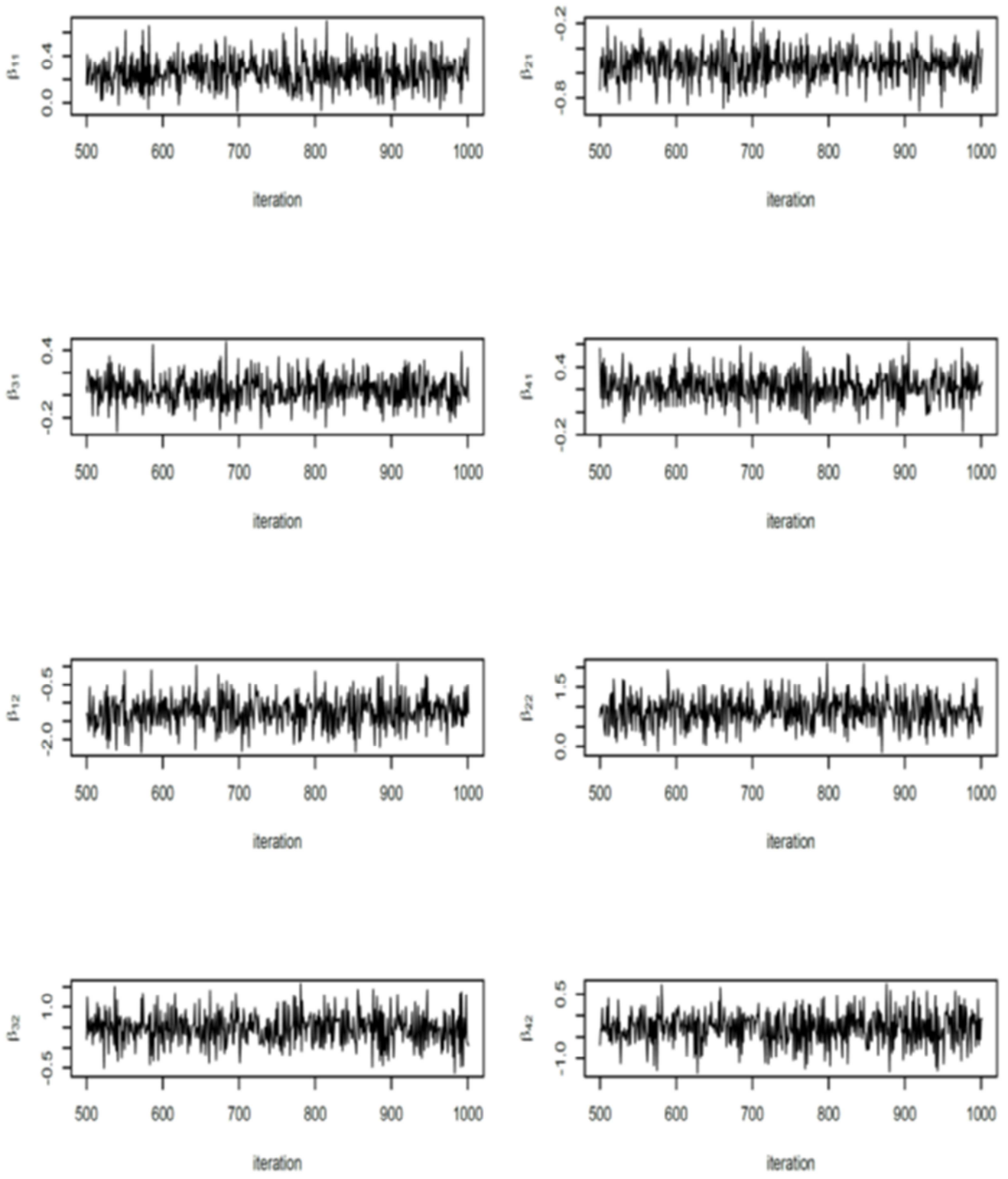

Figure 7. *The sample paths for the coefficients of the first two predictors.*

6. Discussion

In the Bayesian MSVM framework, we need to use the truncated multivariate normal distribution to satisfy the sum to zero constraint on f(.),i.e. $\sum_{r=1}^{k} f_r(\mathrm{x}_i) = 0$ for $i = 1, \cdots, n$, but it sacrifices the efficiency of the computation. In our ca suppose

$$f_r(\mathrm{x}_i) = \mathrm{x}_i \beta_r \tag{30}$$

If we let

$$\mathrm{X} = \begin{pmatrix} \mathrm{x}_1^T \\ \mathrm{x}_2^T \\ \vdots \\ \mathrm{x}_n^T \end{pmatrix}, \mathrm{D} = \begin{pmatrix} \beta_{11} & \beta_{21} & \cdots & \beta_{k1} \\ \beta_{12} & \beta_{21} & \cdots & \beta_{k2} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{1p} & \beta_{21} & \cdots & \beta_{kp} \end{pmatrix} \tag{31}$$

then the sum-to-zero constraint is equivalent to $\mathrm{X}\sum_{r=1}^{k} \beta_r = 0_n$. If the design matrix X is of full rank, then $\sum_{r=1}^{k} \beta_r = 0_p$ or $\mathrm{D}1_k = 0_p$ can guarantee the constraint.

Following [14], one possible solution could be taken the reparameterization procedure below

$$\mathrm{D} = \mathrm{BH} \tag{32}$$

where $\mathrm{H} = \mathrm{I}_k - \frac{1}{k} 1_k 1_k^T$. However, since the matrix H is singular, it is impossible to get the density distribution for the parameters B from the distribution of D by using the density transformation formula. One possible solution for this issue could be solved by looking for a kernel function corresponding to the L_1-normal penalty in (18).

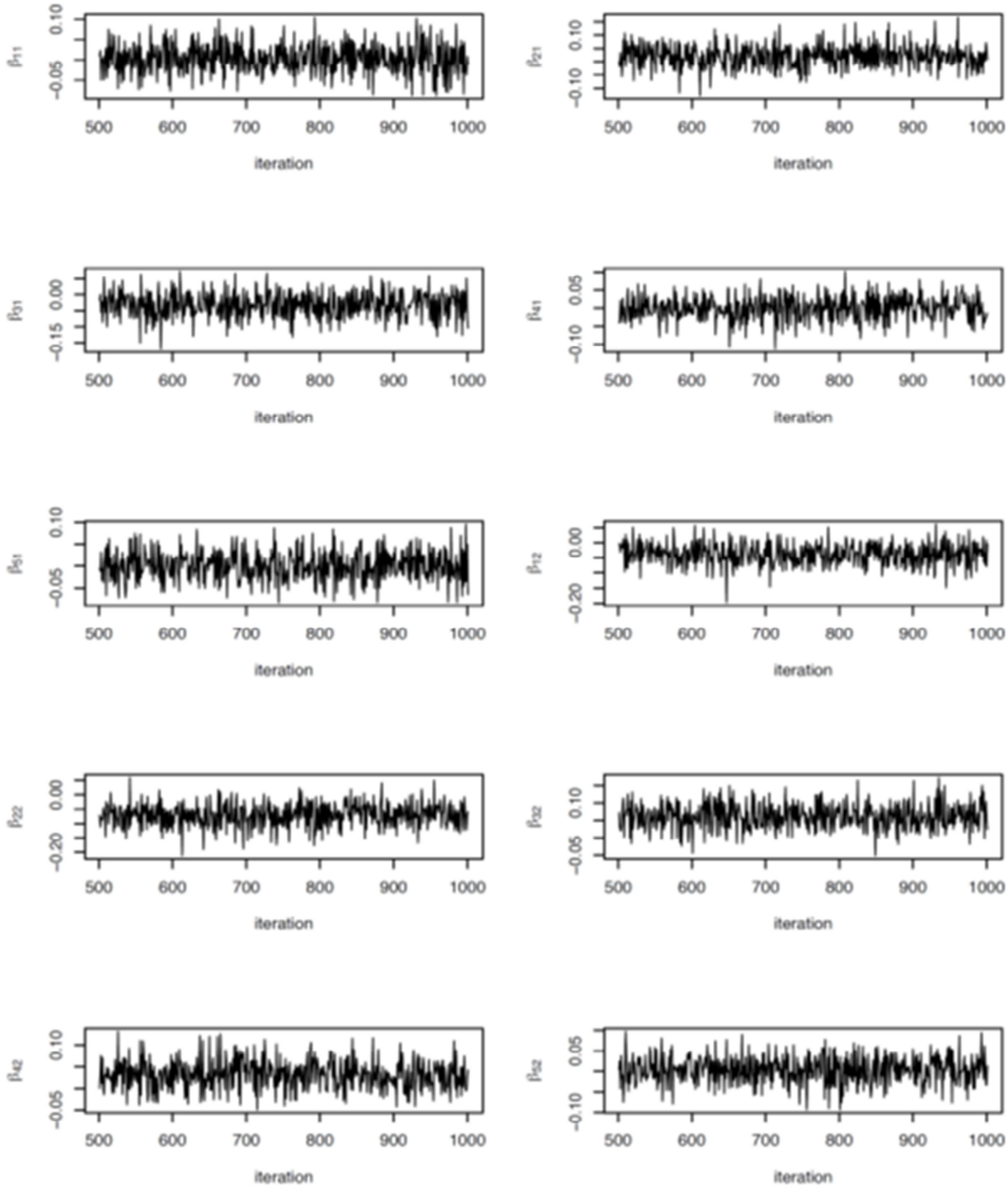

Figure 8. *The sample paths for the coe_cients of the predictors "ITSEX" and "ASBG04".*

7. Conclusion

This paper considered the Bayesian multicategory support vector machine (MSVM). For the problem, a Bayesian framework for MSVM was developed based on stochastic variational inference and inducing points. The proposed Bayesian approach is robust against outliers with advanced accuracy and guaranteed error rate similar to the frequentist formulation of the SVM. The Bayesian method also has the advantage of modeling with high flexibility, automatic parameter tuning, and providing estimates of uncertainty in predictions. The numerical studies suggested that the proposed method has good predication accuracy and works well for practical situations.

References

[1] Boser, B. E., Guyon, I. M., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, 1992.

[2] Vapnik, V. (1998). Statistical learning theory. Wiley, New York.

[3] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2), 121-167.

[4] Nicholas G. Polson and Steven L. Scott (2011). Data Augmentation for Support Vector Machines. *Bayesian Analysis*, 6, 1-24.

[5] Cortes, C., Vapnik, V. (1995) Support-Vector Networks, Machine Learning.

[6] Henao, R., Yuan, X., Carin, L. (2014) Bayesian Nonlinear Support Vector Machines and Discriminative Factor Modeling. Neural Information Processing Systems Conference.

[7] Hoffman, M. D., Blei, D. M., Wang, C., Paisley, J. (2013) Stochastic Variational Inference. Journal of Machine Learning Research, 14, 1303-1347.

[8] Hensman, J., Fusi, N., Lawrence, N. D. (2013) Gaussian processes for big data. Conference on Uncertainty in Artificial Intellegence.

[9] Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2014) Do we need hundreds of classifiers to solve real world classification problems? Journal of Machine Learning Research, 15, 3133-3181.

[10] Mohri, M., Rostamizadeh, A., Talwalkar, A. (2012) Foundations of machine learning. MIT press.

[11] D. F. Andrews and C. L. Mallows (1974). Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society. Series B*, 36, 99-102.

[12] Mike West. (1987). On Scale Mixtures of Normal Distributions. *Biometrika*, 74, 646-648.

[13] Yoonkyung Lee and Zhenhuan Cui (2006). Characterizing the Solution Path of Multicategory Support Vector Machines. *Statistica Sinica* , 16, 391-409.

[14] Zhihua Zhang and Michael I. J ordan (2006). Bayesian Multicategory Support Vector Machines *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence* , 552-559.

[15] Yoonkyung Lee, Yi Lin and Grace Wahba (2004). Multicategory Support Vector Machines Theory and Application to the Classification of Microarray Data and Satellite Radiance Data *Journal of American Statistical Association*, 99, 67-81.