Inaugural EuroConference
Eastern Mediterranean Region
International Biometric Society

Markov Chain Monte Carlo Bayesian Variable Selection for Parametric Survival Models

Ioannis Ntzoufras,* Petros Dellaportas †

Department of Statistics, Athens University of Economics and Business,
76 Patission Street, 10434, Athens, Greece

Jonathan J. Forster ‡

Faculty of Mathematics, University of Southampton, UK

Jan 2001

For postcript file: http://stat-athens.aueb.gr/~jbn/current.htm

2001, Athens, Greece

*e-mail:jbn@aueb.gr
†e-mail:petros@aueb.gr
‡e-mail:jjf@maths.soton.ac.uk

Contents

Introduction: Definitions and Data

- Consider data $\boldsymbol{y}$ divided in
  - survival times (t_i) and
  - incomplete censored times (t_i^*).
- Define as
  - δ_i a censoring time indicator (1 if the exact time of event occurance is available or 0 otherwise),
  - Data: $y_i = \delta_i t_i + (1 - \delta_i) t_i^*$ for all $i = 1, \ldots, n$
  - n is the total number of patients,
  - T_i: the (missing) exact times of event occurance when the information for i subject is censored ($\delta_i = 0$).
- Additional Notation:
  - $f(y_i | \mu_i, \mathbf{r})$ the density function and
  - $S(y_i | \mu_i, \mathbf{r})$ the survival function.
  - μ_i is a shape parameter and $\mathbf{r}$ is a vector of additional parameters.
  - Usually selected distributions are the exponential, Weibull, log-normal, log-logistic and gamma.

Introduction: The Model

- $Y_i \sim f(y_i|\mu_i, \mathbf{r})$ for $i = 1, \ldots, n$,
- with the linear predictor

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \boldsymbol{X}_j \boldsymbol{\beta}_j \qquad (1)$$

  where p is the number of variables/terms in the model.

- Generally $g(\mu_i) = \eta_i$. Usual choice is $g(\mu) = log(\mu)$: $\mu_i = e^{\eta_i}$, for all $i = 1, \ldots, n$; where
  - $\boldsymbol{X}_j$ is the design matrix for j variable or term,
  - $\boldsymbol{\beta}_j$ the parameter vector related to the jth term,
  - η_i is each element of $\boldsymbol{\eta}$ vector.
- The Likelihood is now given by

$$f(\boldsymbol{y}|\boldsymbol{\beta}_m, r) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, \mathbf{r})^{\delta_i} S(y_i|\boldsymbol{\beta}, \mathbf{r})^{1-\delta_i}$$

- The unobserved exact times can be generated by

$$f(T_i|\mu_i, \mathbf{r}) = f(T_i|\mu_i, \mathbf{r}) I(t_i^*, \infty)$$

  where $I(a, b)$ indicates truncation of the distribution at the interval (a, b).

- Finally, by $f(\boldsymbol{\beta}, \boldsymbol{r})$ we denote the prior distribution.

Introduction: Bayesian Model Selection

Bayesian model selection is based on

1. Posterior odds of model m_0 versus model m_1 given by

$$PO_{01} = \frac{f(m_0|\boldsymbol{y})}{f(m_1|\boldsymbol{y})} = \underbrace{\frac{f(\boldsymbol{y}|m_0)}{f(\boldsymbol{y}|m_1)}}_{Bayes\ \ Factor} \times \underbrace{\frac{f(m_0)}{f(m_1)}}_{Prior\ \ Odds}\ .$$

2. Posterior probabilities given by

$$f(m|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|m)f(m)}{\sum\limits_{m_l \in \mathcal{M}} f(\boldsymbol{y}|m_l)f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l,m}\right)^{-1},$$

- m, m_0, m_1 are model indicators,
- $\mathcal{M}$: set of models under consideration.

<u>Why Use MCMC for Model Selection?</u>

1. Automatic after defining the prior distribution,
2. Size of $\mathcal{M}$ is large and hence we cannot explore the model space otherwise,
3. Integrals involved are intractable.
4. Bayesian model averaging is straightforward.

Gibbs Variable Selection

- Introduced by Dellaportas *et al.* (2001)
- We introduce a vector of latent binary indicators $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_p]^T$
  - $\gamma_j = 1$ if j is included in the model
  - $\gamma_j = 0$ if j is excluded from the model
  - Similar approaches: George and McCullogh (1993) and Kuo and Mallick (1998).
- Linear Predictor is changed to

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j \tag{2}$$

  also used by Kuo and Mallick (1998).
- The $f(\boldsymbol{\beta}, \mathbf{r}, \boldsymbol{\gamma})$ is structured as following
  - $f(\boldsymbol{\beta}|\mathbf{r}, \boldsymbol{\gamma})$: Prior of model coefficients given the model,
  - $f(\mathbf{r}|\boldsymbol{\gamma})$: Prior of other parameters given the model,
  - $f(\boldsymbol{\gamma})$: Prior of probability of each model.

- Consider a partition of $\boldsymbol{\beta}$ to
  - $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$: Components of $\boldsymbol{\beta}$ included in the model (with $\gamma_j = 1$),
  - $\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}$: Components of $\boldsymbol{\beta}$ excluded from the model (with $\gamma_j = 0$),

  then the prior $f(\boldsymbol{\beta}|\mathbf{r},\boldsymbol{\gamma})$ may be partitioned into
  - $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\mathbf{r},\boldsymbol{\gamma})$: Prior of model coefficients included in the model,
  - $f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\mathbf{r},\boldsymbol{\gamma})$: Pseudoprior of model coefficients excluded from the model.

- Use multivariate normal prior distribution given by

$$N\left(\boldsymbol{\mu}_{\beta_{(\gamma)}}, c^2\boldsymbol{V}_{(\gamma)}\right). \tag{3}$$

- Pseudoprior does not affect the posterior distribution and is used to assist the MCMC to move efficiently within the model space (It should be close to posterior distribution).

- Each updating step of γ_j in Gibbs Variable Selection is equivalent to Reversible Jump MCMC step; see Ntzoufras (1999).

Gibbs Variable Selection: Sampling steps

The full conditional posterior distributions for the Gibbs variable selection algorithm are given by

1. $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}, \mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{y})$
$$\propto \prod_{i=1}^{n} f(t_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{r}) f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\mathbf{r}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \mathbf{r}, \boldsymbol{\gamma}) \tag{4}$$

2. $f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto$
$$\propto f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \mathbf{r}, \boldsymbol{\gamma}) \tag{5}$$

3. $f(\mathbf{r}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto$
$$\propto \prod_{i=1}^{n} f(t_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{r}) f(\boldsymbol{\beta}|\mathbf{r}, \boldsymbol{\gamma}) f(\mathbf{r}|\boldsymbol{\gamma}) \tag{6}$$

4. $f(t_i|\boldsymbol{\beta}, \mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{y}, \delta_i = 0) \sim f\left(t_i\,|\mu_i, \mathbf{r}\right) I(t_i^*, \infty)$,
$\forall\ i \in \{i = 1, \ldots, n : \delta_i = 0\}$.

5. $f(\gamma_j|\boldsymbol{\beta}, \mathbf{r}, \boldsymbol{\gamma}_{\backslash j}, \boldsymbol{y}) \sim Bernoulli\left(\frac{O_j}{1+O_j}\right)$ with
$$O_j = LR_j \times PR_j \times PrO_j.$$

- LR_j (Likelihood ratio):

$$LR_j = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{r}, \gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{r}, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \tag{7}$$

- PR_j (Prior Ratio):

$$PR_j = \frac{f(\boldsymbol{\beta}|\mathbf{r}, \gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}|\mathbf{r}, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \frac{f(\mathbf{r}|\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\mathbf{r}|\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \tag{8}$$

- PrO_j (Prior Model Odds):

$$PrO_j = \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}. \tag{9}$$

*** NOTE: We may alternatively substitute (4) by

$$\begin{aligned} f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\setminus\boldsymbol{\gamma})}, \mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{y}) & \\ \propto \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{r})^{\delta_i} S(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{r})^{1-\delta_i} \times & \\ \times f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\mathbf{r}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_{(\setminus\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \mathbf{r}, \boldsymbol{\gamma}) & \end{aligned} \tag{10}$$

and skip the generation of the censored survival times given by step 4.

Prior Distributions

- Not able to use flat priors (Lindley's Paradox, 1957).
- Use Normal prior Distribution based on Unit Information (Accounts for information = 1 data point):
  - $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|r, \boldsymbol{\gamma}) = N\left(\mathbf{0}, c^2 \boldsymbol{V}_{(\gamma)}\right)$ with
    $c^2 = n$, $\boldsymbol{V}_{(\boldsymbol{\gamma})} = \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{W}_{(\boldsymbol{\gamma})} \boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1}$
    $\boldsymbol{W}_{(\boldsymbol{\gamma})} = Diag(w_i)$ with w_i given by
    $$w_i = -\left(\frac{\partial l_i{}^2}{\partial \mu_i{}^2}\right)\left(\frac{\partial g(\mu_i)}{\mu_i}\right)^{-2}$$
    where $l_i = \delta_i log\{f(y_i|\mu_i, r)\} + (1-\delta_i) log\{S(y_i|\mu_i, r)\}$.
  - Has similar behaviour to BIC.
  - Disadvantage: Uses data in prior!
- **Objective**: Identify prior distributions based on unit information priors but avoid using data involved.
- If the model is well fitted then $y_i \approx E(y_i)$ and hence we substitute y_i involved in information matrix by their expected values.

The Prior for Weibull Models.

For the weibull models with density

$$f(y|\mu, r) = r y^{r-1} \mu^r e^{-y^r \mu^r}$$

w_i is given by $w_i = r^2 (\mu_i y_i)^r$.

Using $E(y_i) = \Gamma(1 + r^{-1})/\mu_i$ then w_i may be substituted by: $\bar{w} = r^2 \left\{\Gamma(1 + r^{-1})\right\}^r$ for all $i = 1, \ldots, n$.

So we use the prior distribution

$$f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|r, \boldsymbol{\gamma}) = N\left(\mathbf{0}, n\bar{w}^{-1}\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1}\right)$$

$$f(r) = Gamma(10^{-10}, 10^{-10})$$

$$f(\boldsymbol{\gamma}) = Uniform\left(\frac{1}{|\mathcal{M}|}\right)$$

where $|\mathcal{M}|$ is number of models under consideration.

If we use p covariates then $|\mathcal{M}| = 2^p$.

Example 1: Mice Data

Data from Grieve (1987) also analysed by Dellaportas and Smith (1993) and Speigelhalter *et al.* (1996).

- Study on photocarginogenicity with 4 groups, each containing 20 mice.
- We compare the survival times in the three groups:
  - Vehicle control (VC),
  - Test substance group (TS) and
  - Positive control (PC)

  to the baseline group of Irradiated control (IC).
- We consider 3 dummy variables and 8 models.

Details on Results

- We consider 110,000 iterations.
- First 10,000 iterations were discarded as burn-in.
- Convergence of continous and binary parameters is assessed using Geweke diagnostic in CODA software.
- Convergence of posterior Model probabilities is assessed using the batch mean method (25 batches of 4000 iterations).

Model	Post.Prob. (%)	St.Error (±%)
VC	55.4	1.2
VC and TS	27.7	1.5
VC and PC	9.7	0.4
VC, TS and PC	3.4	0.3
Constant	1.7	0.3
PC	1.2	0.4
TS	0.5	0.1
TS and PC	0.4	0.2

Table 1: Posterior Model Probabilities for Mice Dataset.

Model	Post.Prob. (%)	St.Error (±%)
Vehicle control (VC)	96.2	0.5
Test substance (TS)	32.0	1.2
Positive control (PC)	14.7	0.8

Table 2: Posterior Variable Probabilities for Mice Dataset.

Example 2: PBC DATA

Dataset Details

- 276 patients with full recorded data
- Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver
- Conducted between 1974 and 1984
- Source: Fleming, T. and Harrington, D. (1991).
- 17 Variables and 131,072 models.

Details on Results

- Analysis in two steps:
  1. Identify more important variables: Consider all 17 variables and 131,072 models.
  2. Remove variables with posterior probability lower than 0.50 and rerun the algorithm: Consider now only 8 variables and 256 models.

Results

Model	Step 1 P.P. (%)	Step 1 S.E. (±%)	Step 2 P.P. (%)	Step 2 S.E. (±%)
M_0 +ALBUMIN+URINECO+SGOT+PROTHRO	0.92	0.29	10.7	0.8
M_0 +URINECO+SGOT+PROTHRO	0.75	0.34	8.4	1.0
M_0 +ALBUMIN+URINECO+SGOT	0.69	0.23	7.4	0.8
M_0 +URINECO+SGOT	0.65	0.36	7.1	0.7
M_0 +ALBUMIN+URINECO	0.53	0.20	7.1	1.3
M_1 +ALBUMIN+URINECO+SGOT+PROTHRO	0.59	0.20	7.1	0.6
M_0 +ALBUMIN+URINECO +PROTHRO	0.56	0.17	6.9	0.7

Table 3: Posterior Model Probabilities for PBC Data in Step 2. (All other Models have $P.P. < 0.5\%$ in step 1 and $P.P. < 5\%$ in step 2; M_0:AGE+EDEMA+HISTOLOG+SBILIR, M_1:AGE+ HISTOLOG+SBILIR, P.P.:posterior probability, S.E.: MCMC Standard Error based on 25 Batches)

- Step 1: 110,000 iterations.
- Step 2: 510,000 iterations.
- First 10,000 iterations were discarded as burn-in.
- Convergence of continous and binary parameters is assessed using Geweke diagnostic in CODA software.
- Convergence of posterior Model probabilities is assessed using the batch mean method (25 batches of 4,000 iterations or 20,000 iterations).

Variable	Step 1 P.P. (%)	Step 1 S.E. (±%)	Step 2 P.P. (%)	Step 2 S.E. (±%)
TREATMENT	21.3	1.1		
AGE	84.4	16.9	90.8	6.9
SEX	34.0	8.6		
ASCITES	19.1	2.5		
HEPATOME	23.8	3.4		
SPIDERS	21.4	1.5		
EDEMA	79.2	4.6	77.3	2.0
HISTOLOG	90.7	10.1	93.5	2.3
BILIR	94.3	1.2	98.3	0.3
SCHOLES	28.8	3.4		
ALBUMIN	64.7	11.1	68.1	3.5
URINECO	77.2	4.0	86.9	1.8
PHOSPHA	26.9	4.3		
SGOT	60.5	8.1	61.9	3.1
TRIGLYC	15.8	0.6		
PLATELE	19.9	0.7		
PROTHRO	62.6	6.8	63.9	3.0

Table 4: Posterior Variable Probabilities for PBC Data (P.P.:posterior probability, S.E.: MCMC Standard Error based on 25 Batches).

Discussion and Further Research

Further research or structural model characteristics of survival models can be examined using Reversible Jump MCMC algorithm:

- Parametric or non-parametric models?
- Which distrubution is more appropriate?
- Use more general distributions to incorporate special cases (for example generalised F distribution).
- Accelerated or Proportional models (use Lehman family).

Appendix

EXAMPLE 2: VARIABLES UNDER CONSIDERATION IN PBC DATA

Time and Censoring Details:

- TIME :The number of days between registration and the earlier of death, liver transplantation, or study analysis time in July, 1986.
- D :1 if X is time to death, 0 if time to censoring

Explanatory Variables:

1. TREATMENT :Treatment Code (1 = D-penicillamine, -1 = placebo).
2. AGE :Age in years.
3. GENDER :Gender (-1 = male, 1 = female).
4. ASCITES :Presence of ascites (-1 = no, 1 = yes).
5. HEPATOME :Presence of hepatomegaly (-1 = no, 1 = yes).
6. SPIDERS :Presence of spiders (-1 = no, 1 = yes).
7. EDEMA :Presence of edema (-1 = no edema and no diuretic therapy for edema; 0 = edema present for which no diuretic therapy was given, or edema resolved with diuretic therapy; 1 = edema despite diuretic therapy).
8. HISTOLOG :Histologic stage of disease, graded 1, 2, 3, or 4.
9. SBILIR :Serum bilirubin, in mg/dl .
10. SCHOLES :Serum cholesterol, in mg/dl.
11. ALBUMIN :Albumin, in gm/dl.
12. URINECO :Urine copper, in mg/day.
13. PHOSPHA :Alkaline phosphatase, in U/liter.
14. SGOT :SGOT, in U/ml.
15. TRIGLYC :Triglycerides, in mg/dl.
16. PLATELE :Platelet count; coded value is number of platelets per-cubic-milliliter of blood divided by 1000.
17. PROTHRO :Prothrombin time, in seconds.

Variables No 9 - 17 were standardized to avoid large values in MCMC algorithm.

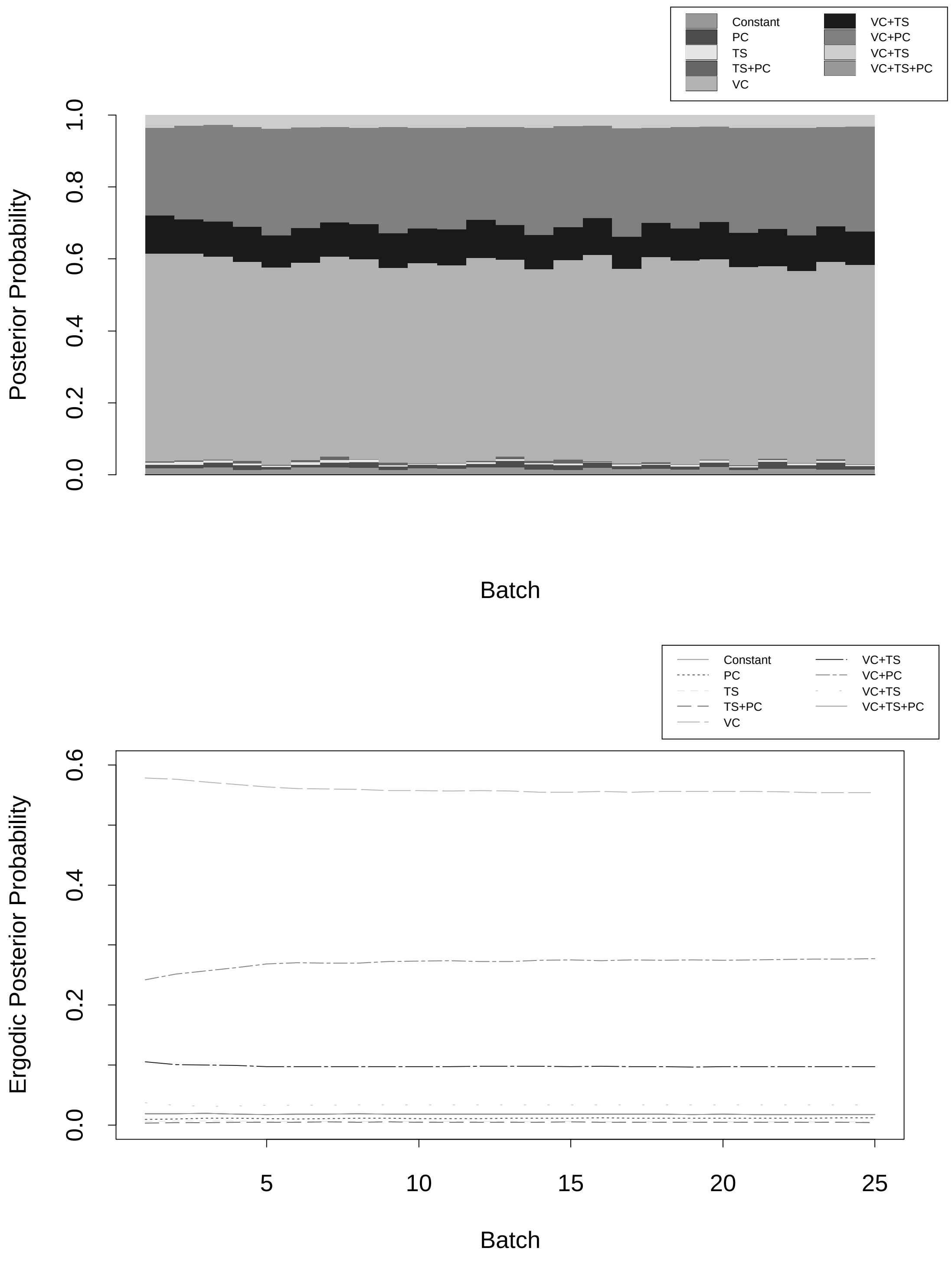


Figure 1: Batch Mean Plots for Example 1.

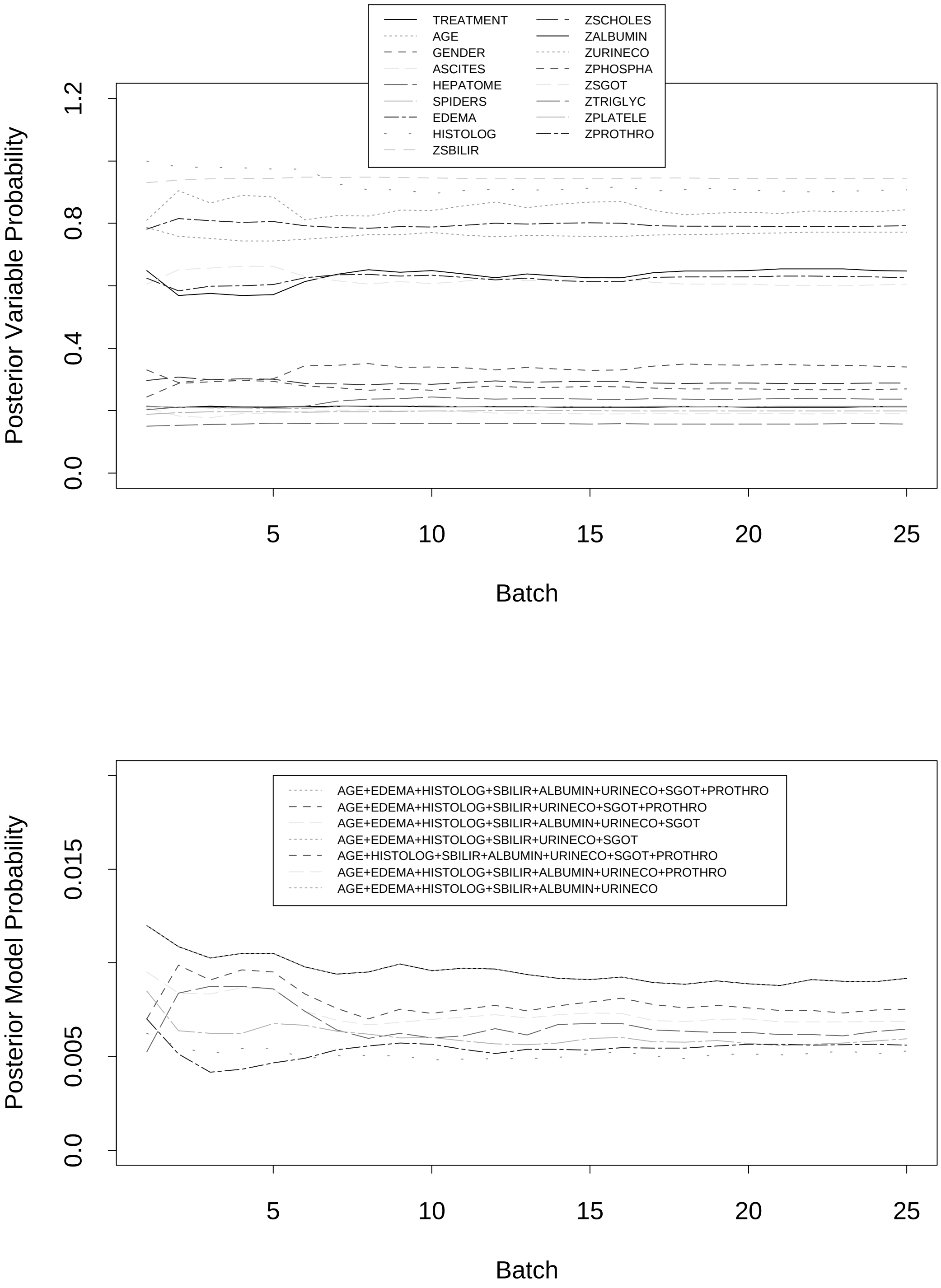
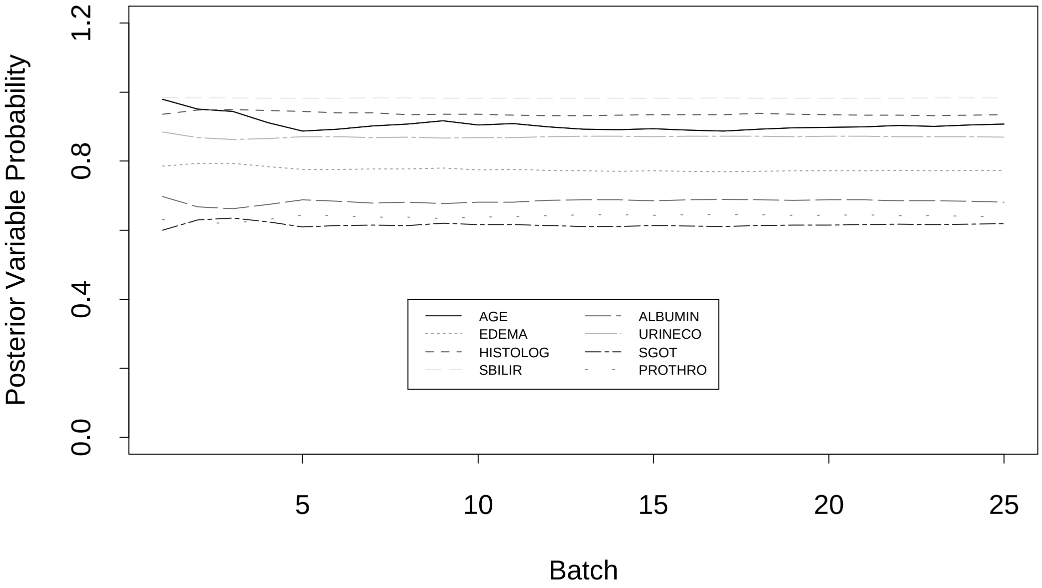


Figure 2: Ergodic Batch Mean Plots for Example 2 (Step 1; 110000 iterations).



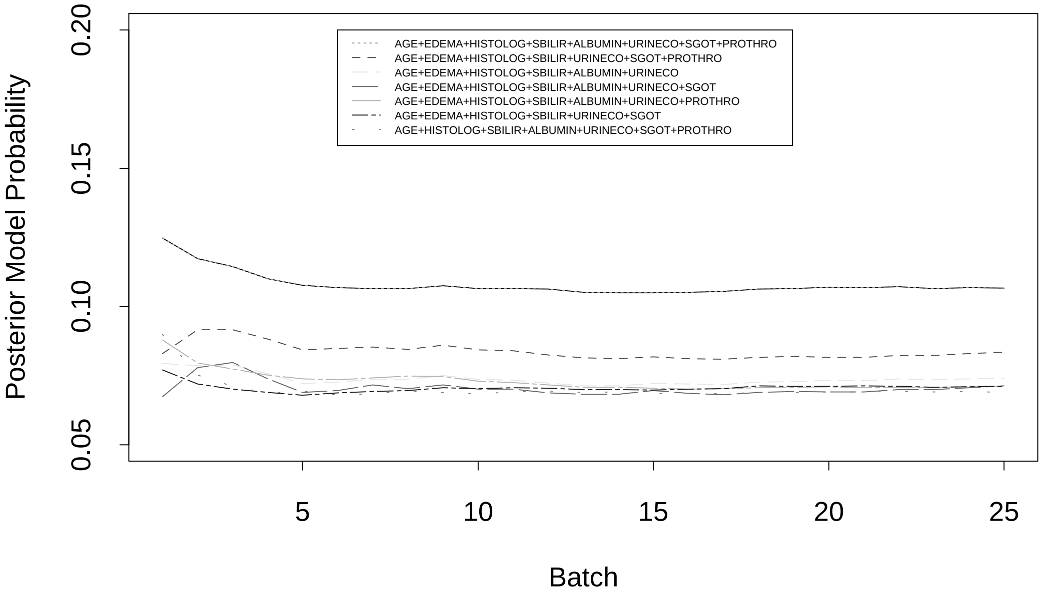


Figure 3: Ergodic Batch Mean Plots for Example 2 ((Step 2; 510000 iterations).

References

[1] Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2001). On Bayesian Model and Variable Selection Using MCMC. *Statistics and Computing*, , (to appear).

[2] Dellaportas, P. and Smith, A.F.M. (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics*, **42**, 443–460.

[3] Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis*. Chichester: John Wiley & Sons.

[4] George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.

[5] Grieve, A.P. (1987). Application of Bayesian Software: Two Examples, *The Statistician*, **36**, 283–288.

[6] Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhyā B*, **60**, 65–81.

[7] Lindley, D.V. (1957). A Statistical Paradox. *Biometrika*, **44**, 187–192.

[8] Ntzoufras, I. (1999). *Aspects of Bayesian Model and Variable Selection Using MCMC. Ph.D. Thesis*, Department of Statistics, Athens University of Economics and Business.

[9] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996b). *BUGS 0.5: Examples Volume 1.* MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.