J. R. Statist. Soc. B (1996)
58, *No.* 2, *pp.* 397–415

Predictive Updating Methods with Application to Bayesian Classification

By RONG CHEN and JUN S. LIU[†]

Texas A&M University, College Station, USA

Stanford University, USA

[Received October 1993. Final revision April 1995]

SUMMARY

We propose algorithms based on random draws from predictive distributions of unknown quantities (missing values, for instance). This procedure can either be iterative, which is a special variation of the Gibbs sampler, or be sequential, which is a variation of sequential imputation. In the latter case one can update the posterior distribution with new observations easily. The methods proposed have intuitive statistical implications and can be generalized to accommodate other Bayesian-like procedures. We display some applications of the method in connection with the Bayesian bootstrap, classification, hierarchical models and selection of variables. In particular, as an application of the method, we present a unified treatment of switching regression models driven by a general binary process, and we develop a Bayesian testing procedure. Some simulations and a real example are used to illustrate the methods proposed.

Keywords: BAYESIAN TESTING; GIBBS SAMPLER; MARKOV CHAIN; ODDS RATIO; SEQUENTIAL IMPUTATION; SWITCHING REGRESSION

1. INTRODUCTION

The idea of multiple imputation arose in the 1970s as a useful tool for dealing with non-response in surveys. The growth of the methodology occurred along with the development of various computational techniques, e.g. the EM algorithm (Dempster *et al.*, 1977), data augmentation (Tanner and Wong, 1987) and methods related to importance sampling. These ideas and techniques led to a widespread recognition of iterative sampling methods in Bayesian analysis (Gelfand and Smith, 1990). Some of these early studies are nicely summarized in Rubin (1987), to which the methods described in this paper are closely related. Here we start with a simple example to introduce the idea of predictive updating for augmenting missing data and updating posterior distributions with new observations. Generalizations to more complicated problems are illustrated in the next few sections.

Suppose that $\mathbf{y}_{\text{obs}} = (y_1, \ldots, y_{n_1})$ is a simple random sample of size n_1 from an unknown population, and let $\boldsymbol{\theta}$ represent all unknown parameters. Furthermore, we assume that there are n_0 additional non-responses $\mathbf{y}_{\text{mis}} = (y_{n_1+1}, \ldots, y_n)$ missing completely at random (Rubin, 1987), where $n = n_0 + n_1$. Obviously, we can ignore the non-responses in making inference for this problem. Here we wish to use this simple example to illustrate the spirit of our proposals.

The key to multiple imputation and data augmentation (Tanner and Wong, 1987)

[†]*Address for correspondence*: Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305-4065, USA.
E-mail: jliu@playfair.stanford.edu

0035–9246/96/58397

is the fact that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ can be expressed as a mixture of completed data posteriors. More precisely, it is recognized that the formula

$$p(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) = \int p(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}})\, p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}})\, \mathrm{d}\mathbf{y}_{\mathrm{mis}}$$

can be realized through Monte Carlo integration. The central task is, therefore, to *impute* multiples of $\mathbf{y}_{\mathrm{mis}} = (y_{n_1+1}, \ldots, y_n)$ by draws from its predictive distribution $p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}})$. Given a draw $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$, it is possible to achieve this imputation by drawing $\mathbf{y}_{\mathrm{mis}}$ from $p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\theta}^*)$. However, sampling from $p(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ is generally difficult, if not impossible. Many recent researches focus on how to circumvent the difficulty. Several methods are available for producing proper multiple imputations in complicated situations.

(a) *Data augmentation*: draw $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}})$; draw $\mathbf{y}_{\mathrm{mis}}$ from $p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\theta})$; then iterate. After many iterations, the $\mathbf{y}_{\mathrm{mis}}$ obtained from this scheme follows from $p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}})$. Tanner and Wong (1987) and Gelfand and Smith (1990) provide more details.

(b) *Iterative predictive update (IPU)*: for $i = n_1 + 1, \ldots, n$, we can iteratively update y_i (i.e. substitute by a new random draw) by drawing from its predictive distribution

$$p(y_i|\mathbf{y}_{\mathrm{obs}}, \mathbf{y}_{\mathrm{mis}[-i]}),$$

i.e. iteratively each y_i is updated by a draw from its predictive distribution conditioned on $\mathbf{y}_{\mathrm{obs}}$ and the current imputed value of $\mathbf{y}_{\mathrm{mis}[-i]}$. In the case of independent and identically distributed (IID) observations from a normal population with unknown mean and variance, under non-informative priors, we only need to draw from a t-distribution with $n-2$ degrees of freedom in each step. This is just a Gibbs sampler applied to draw $\mathbf{y}_{\mathrm{mis}}$ from its target (predictive) distribution $p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}})$.

(c) *Sequential predictive update (SPU)*: for $i = n_1 + 1, \ldots, n$, we sequentially draw y_i^* from its *current* predictive distribution $p(y_i|\mathbf{y}_{\mathrm{obs}}, y_{n_1+1}^*, \ldots, y_{i-1}^*)$. This provides the desired imputation because of the decomposition formula

$$p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}}) = p(y_{n_1+1}|\mathbf{y}_{\mathrm{obs}})\, p(y_{n_1+2}|\mathbf{y}_{\mathrm{obs}}, y_{n_1+1}) \ldots p(y_n|\mathbf{y}_{\mathrm{obs}}, y_{n_1+1}, \ldots, y_{n-1}). \tag{1}$$

Again, in the normal population case, each step is accomplished by drawing from a t-distribution with appropriate degrees of freedom. Compared with the conceptual procedure that draws $\boldsymbol{\theta}$ from its full posterior distribution and then draws $\mathbf{y}_{\mathrm{mis}}$ from $p(\mathbf{y}_{\mathrm{mis}}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\theta})$, the SPU suggests that the parameter uncertainty can be properly reflected by sequentially using predictive distributions. This point will be further illustrated in Section 2.1.

Although the updating schemes are special cases of Gibbs sampler or sequential imputation, they stand out by themselves, for their intuitive statistical interpretations, and for the fact that they can avoid directly modelling the parts of the mechanism that are not of immediate interest. In particular, some nonparametric features can be built in by using the Bayesian bootstrap (BB) (Rubin, 1981) or Dirichlet process modelling (Escobar, 1994; MacEachern, 1994). Furthermore, with some parameters integrated out, we are actually applying the Gibbs sampler to a smaller set of random variables, and therefore the sample autocorrelations and the convergence rate of the resulting sampler are usually better behaved than a full Gibbs sampler (Liu, 1994). In addition, the IPU and SPU methods can avoid unidentifiable problems which occurred in the full Gibbs sampler when dealing with some mixture models and model selection problems.

As a concrete application of the predictive updating schemes, in Section 3 we study a classification problem associated with the switching regression model. This model is a generalization of the mixture density models and can be applied to cases such as threshold autoregressive models, robust regression and pattern recognition. Render and Walker (1984) provided an overview of the EM algorithms for solving the mixture problems. Gibbs sampling techniques have recently been applied to some of the related problems. See, for example, Diebolt and Robert (1994), Lavine and West (1992) and West (1992). More references are provided in Section 3.

The paper is arranged as follows. We display the range of applications of the predictive updating methods in Section 2, which includes their connection with the BB, their use in classification, hierarchical models and selection of variables. Related studies on some special cases can be found in Escobar (1994), Irwin *et al.* (1994) and Lawrence *et al.* (1993). A Bayesian solution of the switching regression problem via predictive updating is provided in Section 3. In Section 4, we propose a procedure for calculating the posterior odds ratio in testing the Markovian dependence in the switching mechanism. In Section 5, we analyse the gross national product (GNP) data of the USA, previously studied by Hamilton (1989) and McCulloch and Tsay (1992), using the techniques described in this paper.

2. PREDICTIVE UPDATING ALGORITHMS

2.1. *Predictive Updating and Bayesian Bootstrap*

The IPU and SPU procedures are closely related to nonparametric Bayes procedures such as Rubin (1981)'s BB and Dirichlet process methods (Escobar, 1994; Liu, 1994, 1995). They are especially useful when only part of the model structure is of direct interest.

Specifically, with $y_1, \ldots, y_{n_1}$ as IID realizations of a random variable Y, one BB replication is generated by drawing $\mathbf{p} = (p_1, \ldots, p_{n_1})$ from a Dirichlet$(1, \ldots, 1)$ distribution. This $\mathbf{p}$ is the vector of probabilities to attach to the data values $y_1, \ldots, y_{n_1}$ in that BB replication. Another way to understand the BB is from the viewpoint of a Dirichlet process. See Lo (1987) for details.

The BB was used for multiple imputations in Rubin (1987). For example, to impute $\mathbf{y}_{\text{mis}} = (y_{n_1+1}, \ldots, y_n)$, we can first draw a BB replication $\mathbf{p}$ to attach to $\mathbf{y}_{\text{obs}}$, and then draw n_0 IID samples from $(y_1, \ldots, y_{n_1})$ with weight $\mathbf{p}$. In the case when $n_0 = 1$, the foregoing procedure is equivalent to drawing directly from $(y_1, \ldots, y_{n_1})$ with equal probability $1/n_1$. This observation immediately gives us a sequential updating procedure which imputes the missing data $\mathbf{y}_{\text{mis}}$ sequentially as follows: for $t = n_1 + 1, \ldots, n$, we let y_t be a simple random draw from the pool

$\{y_1, y_2, \ldots, y_{t-1}\}$. It is easy to show that the resulting $\mathbf{y}_{\text{mis}}$ drawn in this fashion is equivalent to that drawn from the BB procedure by the telescope law (1) and can also be seen as a Polya urn sequence.

This method can be understood as a 'sequential bootstrap' procedure. Although similar to the bootstrap idea, the difference is that each current draw y is reincorporated to produce a future draw. This equivalence of the sequential bootstrap and the BB (or Bayesian prediction) seems to be applicable to more general settings. In addition, the SPU procedure suggests that the extra uncertainty of a future observation (or missing data) due to the unknown parameters can be incorporated *sequentially* through predictive distributions. For example, the 'poor-man's data augmentation' in Tanner (1993) typically underestimates posterior uncertainties and produces improper imputations because it treats the unknown parameter as fixed at its maximum likelihood estimate. However, if the procedure is applied sequentially to each missing observation, which we call the 'poor man's sequential augmentation', proper multiple imputations will be generated.

The IPU procedure can also be modified to draw $\mathbf{y}_{\text{mis}}$ nonparametrically in place of (but equivalent to) the BB procedure. In the above example, for instance, we can iteratively update each y_i by a simple random draw from $\{\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}[-i]}\}$. After many iterations, the joint distribution of $\mathbf{y}_{\text{mis}}$ will converge to that of the samples drawn by the BB procedure. It ties with the Dirichlet process method of Escobar (1994) and MacEachern (1994).

2.2. *Predictive Updating for Classification*

A traditional treatment of classification problems is through the use of finite mixture models, typically normal mixture models. Computationally intensive methods for the maximum likelihood estimates of the parameters of interest have been developed, one of which is the celebrated EM algorithm (Dempster *et al.*, 1977) or its variations. See Everitt and Hand (1981) and Titterington *et al.* (1985) for a review of the area. In applications of these mixture models, however, the classical asymptotic results are no longer trustworthy. Hence, a full Bayesian analysis with appropriate displays of marginal posterior distributions of certain parameters is of great interest. With a data set of moderate size, an explicit analytical solution for a Bayesian analysis is typically formidable. One usually has to employ either some *ad hoc* approximation methods (e.g. Smith and Makov (1978)) or some Monte Carlo methods to complete the analysis. Recent developments in Markov chain Monte Carlo methods for full Bayesian analysis especially help to promote the application of the powerful Bayesian machinery (Diebolt and Robert, 1994; Lavine and West, 1992; West, 1992).

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a set of IID observations coming from a mixture model $\theta_1 p_1(\mathbf{y}, \boldsymbol{\mu}_1) + \ldots + \theta_k p_k(\mathbf{y}, \boldsymbol{\mu}_k)$, where $\theta_1 + \ldots + \theta_k = 1$. Here $\theta_i \geqslant 0$ are the mixing proportions, and the $\boldsymbol{\mu}_i$ are the parameters associated with each distribution.

Let $\mathbf{I} = (I_1, \ldots, I_n)$ be the vector of indicators, i.e. $I_j = l$ if observation $\mathbf{y}_j$ is from class l. If we treat $\mathbf{I}$ as missing data, a natural data augmentation or Gibbs sampler scheme can be implemented as illustrated in Diebolt and Robert (1994) and Lavine and West (1992). However, predictive updating algorithms are more intuitive and can be used to accommodate sequentially arriving data. Specifically, we have

$$\frac{p(I_i = l|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{I}_{[-i]})}{p(I_i = m|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{I}_{[-i]})} = \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\mathbf{I}_{[-i]}, I_i = l)}{p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\mathbf{I}_{[-i]}, I_i = m)} \frac{p(I_i = l|\mathbf{I}_{[-i]})}{p(I_i = m|\mathbf{I}_{[-i]})}. \tag{2}$$

Here

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\mathbf{I}) = \int\left\{\prod_{i=1}^{n} p_{I_i}(\mathbf{y}_i|\boldsymbol{\mu}_{I_i})\right\} \pi(\boldsymbol{\mu}_1) \ldots \pi(\boldsymbol{\mu}_k)\, \mathrm{d}\boldsymbol{\mu}_1 \ldots \mathrm{d}\boldsymbol{\mu}_k.$$

This formula gives us a convenient way for predictive updating, i.e. on the basis of the current classification of the data excluding $\mathbf{y}_j$, we update the 'membership' of the jth observation based on the above probability rule. If the procedure is iterated indefinitely many times, the equilibrium distribution of the above sampler is the posterior distribution of classification $p(\mathbf{I}|\mathbf{y}_1, \ldots, \mathbf{y}_n)$.

Let $\mathbf{I}^{(1)}, \ldots, \mathbf{I}^{(m)}$ be multiple draws from the distribution $p(\mathbf{I}|\mathbf{y}_1, \ldots, \mathbf{y}_n)$ obtained by using some sampler. Now suppose that a new observation $\mathbf{y}_{\text{new}}$ arrives with missing indicator I_{new}. Apparently, the value of I_{new} can be inferred by using the multiple imputed $\mathbf{I}^{(j)}$s and the probability rule (2). A less obvious fact is that the old classifications can be modified on the basis of this new observation, by giving a weight

$$w^{(j)} \propto p(\mathbf{y}_{\text{new}}|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{I}^{(j)})$$

to each $\mathbf{I}^{(j)}$, i.e. each imputation $\mathbf{I}^{(j)}$ is weighted by its ability to predict $\mathbf{y}_{\text{new}}$. This predictive probability is easy to compute in cases such as a normal mixture model with conjugate priors. The general idea of sequential reweighting is explored in depth in Kong *et al.* (1994).

2.3. *Predictive Updating in Hierarchical Models*

Let us consider a typical hierarchical model $\mathbf{y}_i \sim p_i(\mathbf{y}|\boldsymbol{\theta}_i)$ for $i = 1, \ldots, n$, where the $\mathbf{y}_i$ are independent of each other given the $\boldsymbol{\theta}_i$ and the $\boldsymbol{\theta}_i$ are IID from some distribution $G(\boldsymbol{\theta}; \lambda)$. We observe the $\mathbf{y}$s and want to make inference about the $\boldsymbol{\theta}$s and G. In parametric hierarchical models (Morris, 1983), G has a known parametric form. For example, it can be a normal distribution in normal hierarchical models or a beta distribution when we are dealing with proportions and frequencies. In nonparametric cases, however, G is not assumed to have any specific form. Robbins (1955) and Good (1953) treated the unseen species problem under this nonparametric hierarchical framework and obtained the famous 'Robbins–Turing–Good' formula. More recently, Carlin *et al.* (1992) used hierarchical models and the Gibbs sampler to analyse changepoint problems.

Predictive updating methods can be easily applied here to provide proper inference. Suppose now that we consider $\boldsymbol{\theta}_i$ conditioned on $\mathbf{y}_1, \ldots, \mathbf{y}_n$ and $\Theta_{[-i]}$. We can combine two sources of information by the Bayes rule:

$$p(\boldsymbol{\theta}_i|\Theta_{[-i]}, \mathbf{Y}) \propto p(\mathbf{y}_i|\boldsymbol{\theta}_i)\, \mathrm{d}G(\boldsymbol{\theta}_i|\Theta_{[-i]}). \qquad (3)$$

Here $\mathrm{d}G(\boldsymbol{\theta}_i|\Theta_{[-i]})$ represents the corresponding conditional distribution of G. This updating scheme reflects the spirit of hierarchical Bayes modelling that the estimation of a particular $\boldsymbol{\theta}_i$ can be improved by combining information from other similar sources. Formula (3) is easy to implement for parametric models illustrated in Morris (1983) (Gelfand and Smith, 1990). To bring in some nonparametric features, we may wish to update $\mathrm{d}G(\boldsymbol{\theta}_i|\Theta_{[-i]})$ by nonparametric density estimation methods or

Dirichlet process models. Escobar (1994), MacEachern (1984) and Liu (1994, 1995) discussed the use of expression (3) in Bayesian nonparametric hierarchical models where G is assumed to follow a Dirichlet process *a priori*.

2.4. *Predictive Updating for Selection of Variables*

George and McCulloch (1993) nicely illustrated Bayesian methods for the selection of variables. We remark here that the idea of predictive updating can be applied flexibly and has meaningful implications.

Let $\mathbf{Y}$, an $n \times 1$ vector, be the dependent variable, and let $\mathbf{x}_1, \ldots, \mathbf{x}_p$ be a set of potential predictors among which we want to select a subset to fit the 'best' model of the form

$$Y = X_1^* \beta_1^* + \ldots + X_q^* \beta_q^* + \epsilon.$$

Following the notation of George and McCulloch (1993), we introduce the indicator variable $\gamma = (\gamma_1, \ldots, \gamma_p)$, where $\gamma_i = 1$ or $\gamma_i = 0$ according to whether the ith variable $\mathbf{x}_i$ is selected or not. Then

$$\frac{p(\gamma_i = 1|\gamma_{[-i]}, \mathbf{Y}, \mathbf{x}_1, \ldots, \mathbf{x}_p)}{p(\gamma_i = 0|\gamma_{[-i]}, \mathbf{Y}, \mathbf{x}_1, \ldots, \mathbf{x}_p)} = \frac{p(\mathbf{Y}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \gamma_{[-i]}, \gamma_i = 1)}{p(\mathbf{Y}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \gamma_{[-i]}, \gamma_i = 0)} \frac{p(\gamma_i = 1|\gamma_{[-i]}, \mathbf{x})}{p(\gamma_i = 0|\gamma_{[-i]}, \mathbf{x})},$$

In George and McCulloch (1993), the last ratio is just 1 under an 'indifference' prior and is $p_i/(1-p_i)$ under the prior

$$f(\gamma) = \prod p_i^{\gamma_i}(1-p_i)^{1-\gamma_i}.$$

Therefore, the main task involved in an iterative sampler using predictive updating is to compute the predictive probability $p(\mathbf{Y}|\mathbf{x}_1, \ldots, \mathbf{x}_p, \gamma)$. By using suitable priors on the coefficients β_k, this computation is equivalent to integrating out the coefficients in a linear model and can be carried out explicitly (Box and Tiao, 1973). Another benefit of predictive updating in this problem is that it avoids the unidentifiability problem (i.e. $\beta_i = 0$ or $\gamma_i = 0$ cannot be distinguished).

3. CLASSIFICATIONS IN SWITCHING REGRESSION

3.1. *Background*

To illustrate the use of predictive updating schemes, we now provide a detailed analysis of the switching regression problem. Specifically, suppose that observations $(X_1^{\mathrm{T}}, y_1), \ldots, (X_n^{\mathrm{T}}, y_n)$ are from a switching regression model

$$y_i = \begin{cases} X_i^{\mathrm{T}} \beta_1 + \epsilon_i^{(1)}, & \text{if } I_i = 1, \\ X_i^{\mathrm{T}} \beta_2 + \epsilon_i^{(2)}, & \text{if } I_i = 0, \end{cases} \tag{4}$$

where the X_is are p-dimensional column vectors, $\epsilon_i^{(1)} \sim N(0, \sigma_1^2)$ and $\epsilon_i^{(2)} \sim N(0, \sigma_2^2)$ are independent normal errors with possibly different variances. The variables I_i, $i = 1, \ldots, n$, are unobservable indicators that drive the switching mechanism. Our main interest is to identify the values of the indicator. In the later context, we write $\mathbf{I} = (I_1, \ldots, I_n)$.

A brief literature review is as follows. Hosmer (1974) described a case in which the length of halibut of each sex follows different regression models (on age) whereas the sex of fishes cannot be determined cheaply. Quandt and Remsey (1978) used the same model for wage regression. More recently, De Veaux (1989) used the switching regression model to analyse data on musical perception. Shumway and Stoffer (1991) studied a switching linear dynamic model. Those studies are all based on the assumption that the switching indicator is *a priori* exchangeable. A switching mechanism driven by a Markov chain is introduced by Goldfeld and Quandt (1973) in describing housing market disequilibrium. Hamilton (1989) and McCulloch and Tsay (1992), by applying this model to describe the growth in the US GNP, further extended the Markovian structure to time series data.

3.2. *Likelihood and Posterior Distributions*

In the rest of this paper, we use π to denote all the distributions related to the models and parameters. Let $J_i = 1 - I_i$ and $\mathcal{Y} = \{(X_i^{\mathrm{T}}, y_i), i = 1, \ldots, n\}$. The likelihood of model (4) can be written as

$$\pi(\mathcal{Y}|\beta_1, \beta_2, \sigma_1, \sigma_2, \mathbf{I}) \propto \frac{1}{\sigma_1^{n_1}\sigma_2^{n-n_1}} \exp\left[-\sum_{i=1}^{n}\left\{\frac{I_i(y_i - X_i^{\mathrm{T}}\beta_1)^2}{2\sigma_1^2} + \frac{J_i(y_i - X_i^{\mathrm{T}}\beta_2)^2}{2\sigma_2^2}\right\}\right] \tag{5}$$

where

$$n_1 = \sum_{i=1}^{n} I_i.$$

A full Bayesian analysis can be conducted by further assuming prior distributions on the βs, σs and $\mathbf{I}$. We shall show in Section 3.3 that structures of the switching mechanism can be easily modelled via the prior distribution $\pi(\mathbf{I})$.

For a given $\mathbf{I}$, let $n_1 = \Sigma_{i=1}^n I_i$ and

$$S_{xx1} = \sum_{i=1}^{n} I_i X_i X_i^{\mathrm{T}}, \qquad S_{xy1} = \sum_{i=1}^{n} I_i X_i y_i, \qquad \Delta_1 = \sum_{i=1}^{n} I_i y_i^2 - S_{xy1}^{\mathrm{T}} S_{xx1}^{-1} S_{xy1},$$

$$S_{xx2} = \sum_{i=1}^{n} J_i X_i X_i^{\mathrm{T}}, \qquad S_{xy2} = \sum_{i=1}^{n} J_i X_i y_i, \qquad \Delta_2 = \sum_{i=1}^{n} J_i y_i^2 - S_{xy2}^{\mathrm{T}} S_{xx2}^{-1} S_{xy2}.$$

Let p be the number of covariates. The following proposition provides the posterior distributions of the indicator vector $\mathbf{I}$, up to a normalizing constant, in three situations.

Proposition 1. For model (4) with flat priors on β_1 and β_2, and prior $\pi(\mathbf{I})$ on $\mathbf{I}$, the posterior distribution of $\mathbf{I}$ is, by the Bayes theorem, $\pi(\mathbf{I}|\mathcal{Y}) \propto \pi(\mathcal{Y}|\mathbf{I})\,\pi(\mathbf{I})$, provided that $p < \Sigma_{i=1}^n I_i < n - p$, where

(a) if the variances σ_1^2 and σ_2^2 are known then

$$\pi(\mathcal{Y}|\mathbf{I}) \propto \frac{\exp(-\Delta_1/2\sigma_1^2 - \Delta_2/2\sigma_2^2)}{\sigma_1^{n_1-p}\sigma_2^{n-n_1-p}|S_{xx1}S_{xx2}|^{1/2}}, \tag{6}$$

(b) if the variances are unknown but assumed equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and the prior distribution for σ is proportional to $\sigma^{-\alpha}\exp(-\delta/2\sigma^2)$, then

$$\pi(\mathcal{Y}|\mathbf{I}) \propto \frac{1}{(\delta + \Delta_1 + \Delta_2)^{(n-2p-1+\alpha)/2}|S_{xx1}S_{xx2}|^{1/2}}, \tag{7}$$

(c) if the variances are unknown and unequal, and the priors for the variances are proportional to $\sigma_k^{-\alpha_k}\exp(-\delta_k/2\sigma_k^2)$, for $k = 1, 2$, then

$$\pi(\mathcal{Y}|\mathbf{I}) \propto \frac{\Gamma\{(n_1-p+\alpha_1-3)/2\}\,\Gamma\{(n-n_1-p+\alpha_2-3)/2\}}{|S_{xx1}S_{xx2}|^{1/2}}$$

$$\times(\delta_1+\Delta_1)^{-(n_1-p+\alpha_1-1)/2}(\delta_2+\Delta_2)^{-(n-n_1-p+\alpha_2-1)/2}. \tag{8}$$

Proof. First note that

$$\begin{aligned}\sum_{i=1}^{n} I_i(y_i - X_i^{\mathrm{T}}\boldsymbol{\beta}_1)^2 &= \boldsymbol{\beta}_1^{\mathrm{T}}S_{xx1}\boldsymbol{\beta}_1 - S_{xy1}^{\mathrm{T}}\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^{\mathrm{T}}S_{xy1} + S_{yy1}\\ &= (\boldsymbol{\beta}_1 - S_{xx1}^{-1}S_{xy1})^{\mathrm{T}}S_{xx1}(\boldsymbol{\beta}_1 - S_{xx1}^{-1}S_{xy1}) + S_{yy1} - S_{xy1}^{\mathrm{T}}S_{xx1}^{-1}S_{xy1}.\end{aligned}$$

Hence

$$\int_{R^p}\frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} I_i(y_i - X_i^{\mathrm{T}}\boldsymbol{\beta}_1)^2\right\}\mathrm{d}\boldsymbol{\beta}_1 \propto \frac{\exp\{-(S_{yy1} - S_{xy1}^{\mathrm{T}}S_{xx1}^{-1}S_{xy1})/2\sigma^2\}}{n^{1/2}\sigma^{n-p}|S_{xx1}|^{1/2}}. \tag{9}$$

The same calculation can be carried out for $\boldsymbol{\beta}_2$ as well.

Now we prove expression (6). Since flat priors are used on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, the posterior of $\mathbf{I}$, when variances σ_1^2 and σ_2^2 are known, is proportional to

$$\int \pi(\mathcal{Y}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{I})\,\pi(\mathbf{I})\,\mathrm{d}\boldsymbol{\beta}_1\,\mathrm{d}\boldsymbol{\beta}_2.$$

Hence expression (6) can be proved by directly evaluating this integral by using expression (9). To derive expression (8), we integrate out the variance parameters by a simple change of variables: $u = \surd(\delta_1 + \Delta_1)/\sigma_1$ and $v = \surd(\delta_2 + \Delta_2)/\sigma_2$. Hence the problem is reduced to an integral of the beta functions. Expression (7) can be derived similarly. □

The proposition shows that

(a) computations involved in the IPU are easy and
(b) we must be careful when using improper priors.

It is the case for all classification problems that either a proper prior or a restricted space of **I** is required. The priors of β_1 and β_2 are not necessarily flat. When the σs are known, normal distributions are conjugate priors for the βs and can be easily incorporated into the above formulae. When the σs are not known, we assume *a priori* independence between (β_1, σ_1) and (β_2, σ_2) and let $\pi(\beta_k|\sigma_k)$ be $N(b_k, c_k^2\sigma_k^2)$ and

$$\pi(\sigma_k) \propto \sigma_k^{-\alpha_k} \exp(-\delta_k/2\sigma_k^2),$$

for $k = 1, 2$, where the b_k, c_k, α_k and δ_k are hyperparameters and can be input as constants. With those priors, the formulae for $\pi(\mathcal{Y}|\mathbf{I})$ will have the same forms as that in our proposition, with slight modifications. The flat priors are used here only for a simple presentation.

3.3. *Priors on Switching Mechanism*

The switching mechanism can be characterized by introducing another hierarchical parameter Θ which determines the prior distribution of **I**. Several commonly encountered structures can be easily modelled.

3.3.1. *Structure I—completely independent*

In this case, the I_is are assumed independent Bernoulli random variables, being 1 with probability θ_i and 0 with probability $1 - \theta_i$. This situation occurs when the population changes from observation to observation. It is easily seen that the resulting $\pi(\mathbf{I}|\Theta)$ is proportional to

$$\prod_{i=1}^{n} \theta_i^{I_i}(1 - \theta_i)^{J_i}.$$

If the θ_is are independent and $\theta_i \sim \text{beta}(a_i, b_i)$ then

$$\pi(\mathbf{I}) \propto \prod_{i=1}^{n} a_i^{I_i} b_i^{J_i}.$$

3.3.2. *Structure II—exchangeable*

When all the observations are simple random draws from a mixture of two subpopulations, a reasonable and commonly used mechanism is the exchangeable structure in which the I_i are assumed to be IID Bernoulli random variables with a common unknown parameter θ. Hence, $\pi(\mathbf{I}|\Theta) = \theta^{n_1}(1 - \theta)^{n-n_1}$, where $n_1 = \Sigma_{i=1}^{n} I_i$. If a conjugate prior distribution $\theta \sim \text{beta}(a, b)$ is used, then the resulting prior distribution of **I**, after integrating θ out, is

$$\pi(\mathbf{I}) \propto \Gamma(n_1 + a)\,\Gamma(n - n_1 + b).$$

In the switching regression literature, this structure has most commonly been used. See Hosmer (1974). Quandt and Remsey (1978), Shumway and Stoffer (1991) and De Veaux (1989) for their methods.

3.3.3. *Structure III—Markovian dependent*

When observations come in sequentially, it is reasonable to assume that the switching mechanism is time dependent and the indicator variables follow a binary process. The simplest binary process is the first-order two-state Markov chain, which has been used by Hamilton (1989) and McCulloch and Tsay (1992) in a time series model.

Assume that the transition probabilities for the indicator Markov chain are

$$\pi(I_{i+1} = 1 | I_i = 1, \Theta) = \theta_{11}$$

and

$$\pi(I_{i+1} = 0 | I_i = 0, \Theta) = \theta_{00},$$

with initial state $\pi(I_1 = 1|\Theta) = \delta$; then

$$\pi(\mathbf{I}|\Theta) = \theta_{11}^{n_{11}} (1 - \theta_{11})^{n_{10}} \theta_{00}^{n_{00}} (1 - \theta_{00})^{n_{01}} \delta^{I_i} (1 - \delta)^{J_i},$$

where n_{ij} is the number of transitions from i to j among the sequence $I_1, \ldots, I_n$. Computationally, $n_{11} = \Sigma_{i=1}^{n-1} I_i I_{i+1}$, $n_{10} = \Sigma_{i=1}^{n-1} I_i J_{i+1}$, $n_{00} = \Sigma_{i=1}^{n-1} J_i J_{i+1}$ and $n_{01} = \Sigma_{i=1}^{n-1} J_i I_{i+1}$, where $J_i = 1 - I_i$. Suppose that the prior distribution of $(\theta_{00}, \theta_{11})$ is a product of independent beta distributions with parameters (a_0, b_0) and (a_1, b_1), and let δ follow a uniform distribution on [0, 1]. Then the marginal prior distribution of **I** is

$$\pi(\mathbf{I}) \propto \frac{\Gamma(n_{11} + a_1)\,\Gamma(n_{10} + b_1)\,\Gamma(n_{00} + a_0)\,\Gamma(n_{01} + b_0)}{\Gamma(n_{11} + n_{10} + a_1 + b_1)\,\Gamma(n_{00} + n_{01} + a_0 + b_0)}.$$

The same approach can be easily extended to a kth-order Markov chain.

3.3.4. *Structure IV—balance switching models*

Let us consider a system where the probability of switching depends on the time length since its previous switch, i.e.

$$\pi(I_i = 1 | I_{i-1} = 1, \ldots, I_{i-k} = 1, I_{i-k-1} = 0) = \theta_{k1}$$

and

$$\pi(I_i = 0 | I_{i-1} = 0, \ldots, I_{i-k} = 0, I_{i-k-1} = 1) = \theta_{k0}$$

for $k = 0, 1, \ldots, n$. We call the system the *balance switching model* if the θ_{ki} decreases as k increases. This structure might be particularly interesting to economists and engineers since economic conditions and engineering systems have memory and tend to change after a stable period.

One way to achieve balance is to introduce independent beta prior distributions for the θ_ks, i.e. *a priori* we assume $\pi(\theta_{k1}, \theta_{k0}) \sim \text{beta}(a_{k1}, b_{k1})\,\text{beta}(a_{k0}, b_{k0})$, and let the a_k decrease and the b_k increase. Specifically, since

$$\pi(\mathbf{I}|\Theta) = \prod_{i=1}^{n} \pi(I_i | I_1, \ldots, I_{i-1}, \Theta) = \prod_{k=0}^{n} \theta_{k1}^{n_{k1}} (1 - \theta_{k1})^{m_{k1}} \theta_{k0}^{n_{k0}} (1 - \theta_{k0})^{m_{k0}}$$

where n_{k1} is the number of sequence (01. . .11) with total $k+1$ 1s, m_{k1} is the number of sequence (01. . .10) with total k 1s, n_{k0} is the number of sequence (10. . .00) with total $k+1$ 0s and m_{k1} is the number of sequence (10. . .01) with total k 0s, we have

$$\pi(\mathbf{I}) \propto \prod_{i=0}^{n} \frac{\Gamma(n_{k1}+a_{k1})\,\Gamma(m_{k1}+b_{k1})\,\Gamma(n_{k0}+a_{k0})\,\Gamma(m_{k0}+b_{k0})}{\Gamma(n_{k1}+m_{k1}+a_{k1}+b_{k1})\,\Gamma(n_{k0}+m_{k0}+a_{k0}+b_{k0})}.$$

This strategy will be implemented in the GNP example of Section 5. Another way to achieve balance might be to employ parametric functions $\theta_{k1}=g_1(k, \mathcal{A}_1)$ and $\theta_{k0}=g_0(k, \mathcal{A}_0)$ where $\mathcal{A}_1$ and $\mathcal{A}_0$ are hyperparameters and the functions g_1 and g_0 decrease with k. With this structure, the parameter $\boldsymbol{\Theta}$ can no longer be integrated out easily in the posterior distribution.

3.3.5. *Other structures*

We mention two more structures that will not be studied in detail here. Tong (1983) introduced the threshold autoregressive (TAR) model of the form

$$x_t = \begin{cases} \phi_1^{(1)}x_{t-1} + \ldots + \phi_p^{(1)}x_{t-p} + \epsilon_t^{(1)} & \text{if } x_{t-d} < c, \\ \phi_1^{(2)}x_{t-1} + \ldots + \phi_p^{(2)}x_{t-p} + \epsilon_t^{(2)} & \text{if } x_{t-d} \geqslant c. \end{cases}$$

If we let $X_i=(x_{i-1}, \ldots, x_{i-p})$ and let $I_i(c)=1$ when $x_{i-d}<c$, and let $I_i(c)=0$ otherwise, then the TAR model fits nicely into the switching regression framework, conditioned on the first p observations. Since $\mathbf{I}$ is completely determined by c, the problem of classification is reduced to that of determining c. With a flat prior on c, the posterior distribution of c is proportional to $\pi(\mathcal{Y}|\mathbf{I})$ in expressions (6)–(8) with all the I_is replaced by the $I_i(c)$ defined above. In this case the computation can easily be done. A similar Bayesian model has been studied by Geweke and Terui (1991) using numerical integration procedures.

The switching mechanism can also be driven by a logistic regression model. In a recent study by Rubin and Wu (personal communication), I_i is assumed to follow a Bernoulli distribution with parameter θ_i, and $\text{logit}(\theta_i)=Z_i\gamma$, where Z_i can be X_i, a subcomponent of X_i or some exogenous variable. A multivariate normal prior, for example, can be imposed on the logistic coefficient γ. This problem was studied by Rubin and Wu by using an EM-type algorithm. A Gibbs sampling approach is easily available.

3.4. *Implementation*

We applied the IPU procedure to produce random draws from the posterior distribution of the indicator vector $\mathbf{I}$. Specifically, for a fixed i, the conditional posterior probability $\pi(I_i=1|\mathbf{I}_{[-i]}, \mathcal{Y})$ is computed. We shall show that this step is easy because of a recursive formula. Then a Bernoulli random variable is generated with this probability and the I_i is updated accordingly. With trivial starting positions and given hyperparameters of the prior distributions, the procedure runs through $i=1, \ldots, n$ to complete one cycle of IPU iteration. In practice, $M+N$ iterations are needed, of which the first M iterations are discarded and the last N iterations are saved as posterior draws from the true posterior distribution of $\mathbf{I}$.

For each update only one observation's membership is under consideration. The computation of $\pi(\mathcal{Y}|I_i, \mathbf{I}_{[-i]})$ can be efficiently done by using the following recursive formulae which do not require the computation of determinants and matrix inversions. For a given $\mathbf{I}^0$, let S^0_{xx1} and S^0_{xx2} be the corresponding sample matrices as defined in Section 3.2, and let P^0_{xx1} and P^0_{xx2} be their respective inverses. Let $I^0_i = \omega$ with ω being either 1 or 0, and let $\mathbf{I}^* = (I_i = 1 - \omega, \mathbf{I}^0_{[-i]})$. To compute $\pi(\mathcal{Y}|\mathbf{I}^*)$, we only need the ratio $\pi(\mathcal{Y}|\mathbf{I}^*)/\pi(\mathcal{Y}|\mathbf{I}^0)$, which can be viewed as the odds of removing the ith observation (X_i, y_i) from its current group (regression line) and adding it to the opposite group. The following relationships hold after I^0_i has been changed from ω to $1 - \omega$:

$$P^*_{xx1} = P^0_{xx1} - \frac{P^0_{xx1} X_i X_i^{\mathrm{T}} P^0_{xx1}}{X_i^{\mathrm{T}} P^0_{xx1} X_i + (-1)^{\omega}},$$

$$P^*_{xx2} = P^0_{xx1} - \frac{P^0_{xx2} X_i X_i^{\mathrm{T}} P^0_{xx2}}{X_i^{\mathrm{T}} P^0_{xx2} X_i - (-1)^{\omega}},$$

and for the determinants

$$|S^*_{xx1}| = |S^0_{xx1}||X_i^{\mathrm{T}} P^0_{xx1} X_i + (-1)^{\omega}|,$$

$$|S^*_{xx2}| = |S^0_{xx2}||X_i^{\mathrm{T}} P^0_{xx1} X_i - (-1)^{\omega}|.$$

These equations can be derived by using standard matrix algebra and the fact that $S^*_{xx1} = S^0_{xx1} - (-1)^{\omega} X_i X_i^{\mathrm{T}}$ and $S^*_{xx2} = S^0_{xx2} + (-1)^{\omega} X_i X_i^{\mathrm{T}}$ when the ith observation (X_i^{T}, y_i) has been removed from its current group and added to its opposite group. Similar computations can be found in Chen and Tsay (1993).

3.5. *Simulation Study*

To illustrate the performance of the proposed procedure, 400 observations were simulated from the model

$$y_i = \begin{cases} 0.7x_{i1} - 0.3x_{i2} + \epsilon_i & \text{if } I_i = 1, \\ 1 - 0.7x_{i1} + 0.8x_{i2} + \epsilon_i & \text{if } I_i = 0 \end{cases}$$

Where x_{i1} and x_{i2} were generated from a uniform distribution on $[-4, 4]$ and the ϵs were normal random variables with mean 0 and standard deviation 0.5. The indicator series I_i were generated from a Markov chain with switching probability 0.05, i.e. $\pi(I_i = 1|I_{i-1} = 1) = \pi(I_i = 0|I_{i-1} = 0) = 0.95$. We computed the posterior distributions of $\mathbf{I}$ under both the exchangeable switching model and the Markov-chain-driven switching model. In both cases, 1000 iterations of the IPU were carried out and the first 500 were discarded. For the exchangeable model, a beta(1, 1) distribution was used as the prior density for the common θ. For the Markovian model, a product of two beta(1, 1) densities was used as the prior for $(\theta_{11}, \theta_{00})$. Fig. 1 shows the corresponding posterior probabilities $\pi(I_i = 1|\mathcal{Y})$. The dots are the 'true' indicators and the lines are the posterior probabilities. We can see clearly that the

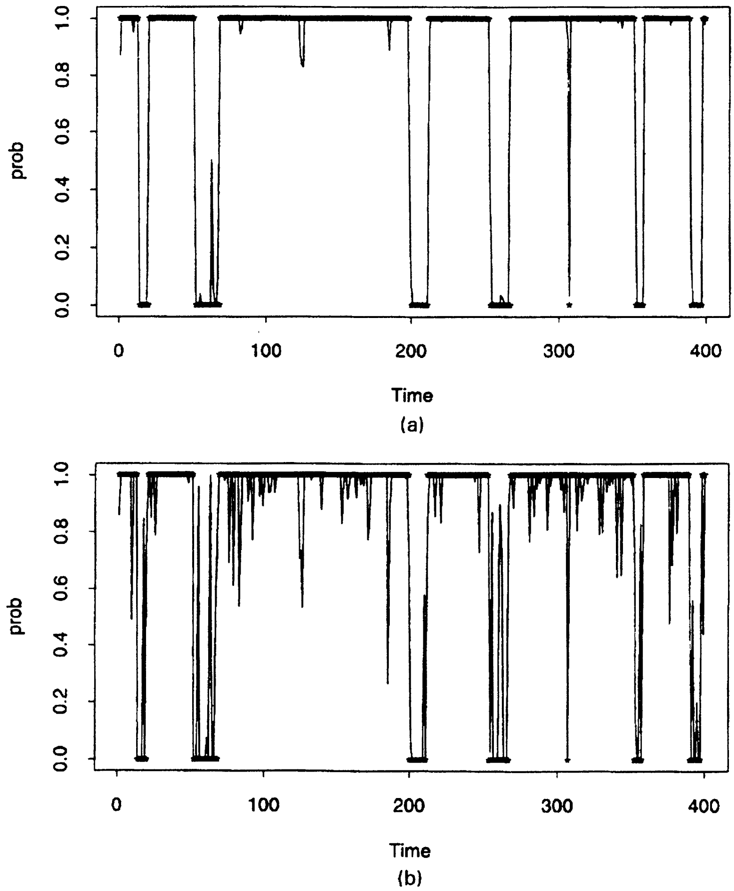

Fig. 1. Posteriors for I_i for a data set generated from structure III (*, true indicator; –, posterior probability): (a) Markovian and (b) exchangeable model

Markovian structure, which is the underlying true structure, works much better. Another example with the I_is IID from a Bernoulli(0.5) distribution was simulated. We carried out the same computations as in the previous example with the same prior distributions. Since states 0 and 1 are highly mixed, a figure like Fig. 1 is not suitable. In Fig. 2, we plot the posterior probability of $\pi(I_i = I_i^* | \mathcal{Y})$ where I_i^*s are the true state of the process generated. The classification via an exchangeable model was as good as that via a Markovian model.

3.6. *Discussion*

It is noted that an unidentifiability problem is present since state 1 and state 0 are exchangeable, i.e. marginally I_i is equally likely to be 1 or 0, which reveals a feature of bimodality for the posterior distribution of $\mathbf{I}$. Such a problem can be avoided in two ways. The first is to choose the optimal classification $\hat{\mathbf{I}}$ as the maximum *a posteriori*, i.e. the classification that maximizes $\pi(\mathbf{I} | \mathcal{Y})$. The parameters β_i and σ_i can be estimated accordingly. A second way is to confine ourselves on one mode of the distribution $\pi(\mathbf{I} | \mathcal{Y})$ by imposing certain constraints on the (hidden) regression

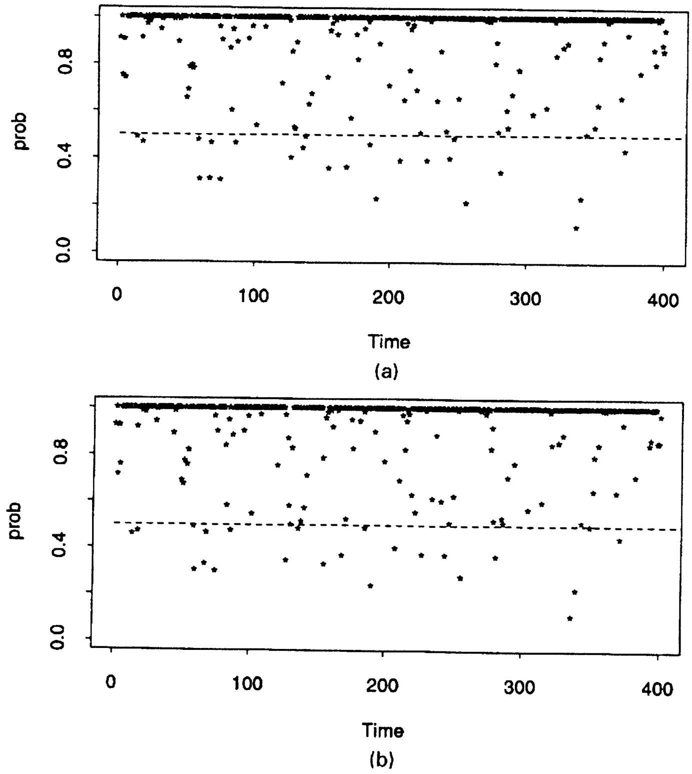

Fig. 2. Posteriors for I_i for a data set generated from structure II (*, posterior probability $\pi(I_i = I_i^* \mid \mathcal{Y})$ where I_i^* are the true states of the generated process): (a) Markovian and (b) exchangeable model

parameters. More precisely, the **I**s sampled from $\pi(\mathbf{I} \mid \mathcal{Y})$ by the IPU are classified as those with β_1 smaller than β_2 and those with β_1 greater than β_2. Then the two groups correspond to the two modes of $\pi(\mathbf{I} \mid \mathcal{Y})$.

Another *ad hoc* method is to run a restricted IPU with some constraints such as $\beta_1 > \beta_2 + \delta$, i.e. the sampler only accepts those draws satisfying the constraints. This amounts to putting a vague prior distribution on **I** and the resulting limiting distribution will not be the same as $\pi(\mathbf{I} \mid \mathcal{Y})$. We tested all the above methods. They seem not to make any practical differences to the test data sets that we tried.

Our formulation can be easily generalized to multilevel Bayes models with the Is taking values in more than two levels. In addition, sensible prior information on various parameters can easily be incorporated, and hierarchical structures on the switching mechanism can be imposed flexibly. The nuisance parameters are effectively eliminated and thus the 'core' part of the likelihood function, $\pi(\mathcal{Y} \mid \mathbf{I})$, is extracted, with a recursive formula for updating. The 'three-scheme' theorem of Liu (1994) shows that, by eliminating all the nuisance parts (even if these parts can be computed and drawn cheaply), the IPU procedure applied to the core part usually converges faster and has lower autocorrelations between samples than a straightforward Gibbs sampler applied directly to the joint posterior distribution of all the parameters. Therefore, our treatment is computationally more efficient than that of

McCulloch and Tsay (1992), even when we are primarily interested in the regression coefficients.

4. TESTING EXCHANGEABILITY AGAINST MARKOVIAN DEPENDENCE

Recently, hidden Markov models similar to structure III in Section 3.3 have been developed and extensively studied in speech recognition (Rabiner, 1989) and other fields. It is often desirable to test the necessity of this structure for a given set of data, i.e. we would be more willing to use an exchangeable model if a Markovian dependence model does not suggest itself.

Following the notation in structure III of Section 3.3, we let $\theta_{01} = 1 - \theta_{00}$. We are interested in testing H_0: $\theta_{01} = \theta_{11}$ *versus* H_1: $\theta_{01} \neq \theta_{11}$. Note that the null hypothesis corresponds to the exchangeable structure. If we impose a prior of the form

$$\pi(\theta_{01}, \theta_{11}) \propto \alpha\, \pi_0(\theta_{01})\, \delta(\theta_{01} = \theta_{11}) + (1 - \alpha)\, \pi_1(\theta_{01}, \theta_{11}),$$

where $\alpha/(1 - \alpha)$ is called the prior odds ratio, then the *posterior odds ratio*

$$r = \frac{\alpha\, \pi_0(\mathcal{Y})}{(1 - \alpha)\, \pi_1(\mathcal{Y})}$$

is of interest to Bayesians. Since r cannot be computed explicitly in our problem, a Monte Carlo method is needed. Liu (1994) noticed that a standard Gibbs sampler is not directly applicable for such a computation because of the degeneracy of H_0 with respect to H_1. Thus a type of IPU as in Liu (1994) with the θ integrated out is needed to resolve the difficulty.

Suppose that conjugate priors $\pi_0(\theta_{01}) = \text{beta}(a, b)$ and $\pi_1(\theta_{01}, \theta_{11}) = \text{beta}(a_0, b_0)\,\text{beta}(a_1, b_1)$ are used for Θ and let α be 0.5. Then the posterior distribution of $\mathbf{I}$ is

$$\pi(\mathbf{I} \mid \mathcal{Y}) \propto \pi(\mathcal{Y} \mid \mathbf{I})\{A_0(\mathbf{I}) + A_1(\mathbf{I})\}/2,$$

where

$$A_0(\mathbf{I}) = \frac{\Gamma(n_{11} + n_{01} + a)\, \Gamma(n_{10} + n_{00} + b)}{\Gamma(n_{11} + n_{01} + n_{10} + n_{00} + a + b)}\, \frac{\Gamma(a + b)}{\Gamma(a)\, \Gamma(b)},$$

$$A_1(\mathbf{I}) = \frac{\Gamma(n_{11} + a_1)\, \Gamma(n_{10} + b_1)\, \Gamma(n_{00} + b_0)\, \Gamma(n_{01} + a_0)}{\Gamma(n_{11} + n_{10} + a_1 + b_1)\, \Gamma(n_{00} + n_{01} + a_0 + b_0)}\, \frac{\Gamma(a_0 + b_0)\, \Gamma(a_1 + b_1)}{\Gamma(a_0)\, \Gamma(b_0)\, \Gamma(a_1)\, \Gamma(b_1)}.$$

Note that the posterior distribution using prior π_0 on Θ under H_0 is $\pi_0(\mathbf{I} \mid \mathcal{Y}) \propto A_0(\mathbf{I})\, \pi(\mathcal{Y} \mid \mathbf{I})$ and that using prior π_1 on Θ is $\pi_1(\mathbf{I} \mid \mathcal{Y}) \propto A_1(\mathbf{I})\, \pi(\mathcal{Y} \mid \mathbf{I})$.

The IPU procedure is then applied to draw from $\pi_1(\mathbf{I} \mid \mathcal{Y})$. Let $\mathbf{I}^{(1)}, \ldots, \mathbf{I}^{(N)}$ be samples obtained from such a geometric mixing sampler. For each sample $\mathbf{I}^{(k)}$, the value $A_0(\mathbf{I}^{(k)})/A_1(\mathbf{I}^{(k)})$ is computed. Then, by Liu (1994), the odds ratio can be approximated by the ergodic average:

$$r \approx \frac{1}{N}\left\{\frac{A_0(\mathbf{I}^{(1)})}{A_1(\mathbf{I}^{(1)})}+\ldots+\frac{A_0(\mathbf{I}^{(N)})}{A_1(\mathbf{I}^{(N)})}\right\}.$$

By similar arguments, if we sample $\mathbf{I}$ from $\pi_0(\mathbf{I})$ (the exchangeable model) and compute the ratio $A_1(\mathbf{I})/A_0(\mathbf{I})$, their average converges to $1/r$ almost surely.

The above test via the IPU is closely related to the importance sampling idea. We found that sampling $\mathbf{I}$ from the larger model, i.e. $\pi_1(\mathbf{I})$ in our case, usually resulted in a smaller coefficient of variation of the ratio than that from the smaller model $\pi_0(\mathbf{I})$. This suggests that $\pi_1(\mathbf{I})$ is more efficient.

5. REAL EXAMPLE

Assuming that the economy can be identified as in 'contraction' or 'expansion' states, Hamilton (1989) and McCulloch and Tsay (1992) studied the quarterly real GNP of the USA (from the first quarter of 1947 to the first quarter of 1991) using different autoregressive structures for the two states. The data are in billions of 1982 dollars, seasonally adjusted. The transformation $y_t = \log(x_t/x_{t-1})$ was taken. The data are shown in Fig. 3(a). For detailed information, see McCulloch and Tsay (1992). Here we repeat their analysis by using the methods proposed in this paper.

We applied the method in Section 4 to test whether a (hidden) Markov switching mechanism, which is the basic assumption of the analysis of Hamilton (1989) and McCulloch and Tsay (1992), is necessary for the data set. When flat priors are used for Θ, the posterior odds ratio r was estimated as 1.67 and 1.57 in two Monte Carlo computations each with 6000 IPU runs (the first 1000 runs were discarded). The evidence of Markovian dependence was not at all strong. This partially explains why very strong priors on the indicator are needed to enforce a Markovian structure, as in McCulloch and Tsay (1992), who used $\pi_1(\theta_{01}, \theta_{11}) = \text{beta}(5, 45)\,\text{beta}(45, 5)$. Interestingly, when their choice of π_1 was used against a flat π_0, we obtained a much smaller r. Two Monte Carlo computations estimated r as 1.47×10^{-5} and 4.79×10^{-5} each with 6000 IPU runs.

With $I=1$ corresponding to the contraction state, Fig. 3(b) shows the posterior probabilities $\pi(I_i = 1 \mid \mathcal{Y})$ from using a switching AR(2) model with a Markov chain switching mechanism and a product of beta(45, 5) distributions as the *a priori* distribution for $(\theta_{11}, \theta_{00})$. This choice is similar to that of Hamilton (1989) and McCulloch and Tsay (1992). Our results were obtained by using 6000 IPU iterations with the first 1000 runs discarded. Fig. 3(c) uses the balanced switching model with prior $\text{beta}(50-2k,\ 1+2k)$ for θ_{k1} and θ_{k0}, $k = 0, \ldots, 24$, $\text{beta}(2, 45)$ for $k = 25, \ldots, 30$ and $\theta_{k1} = \theta_{30,1}$, and $\theta_{k0} = \theta_{30,0}$ for $k > 30$. The results are similar to those of Hamilton (1989) and McCulloch and Tsay (1992), though the posterior probabilities of contraction that we obtained are generally higher. In Figs 3(b) and 3(c), we also marked the time when the economy reached its peaks (at the bottom of the figures) and troughs (at the top of the figures) of the business cycle. The dates of these peaks and troughs are published by the National Bureau of Economic Research, Inc. We can see that these dates are well fitted in the contraction and expansion picture that we obtained. It is also seen that the date for a trough is usually earlier than the time that the posterior probability of a contraction hits its local peak

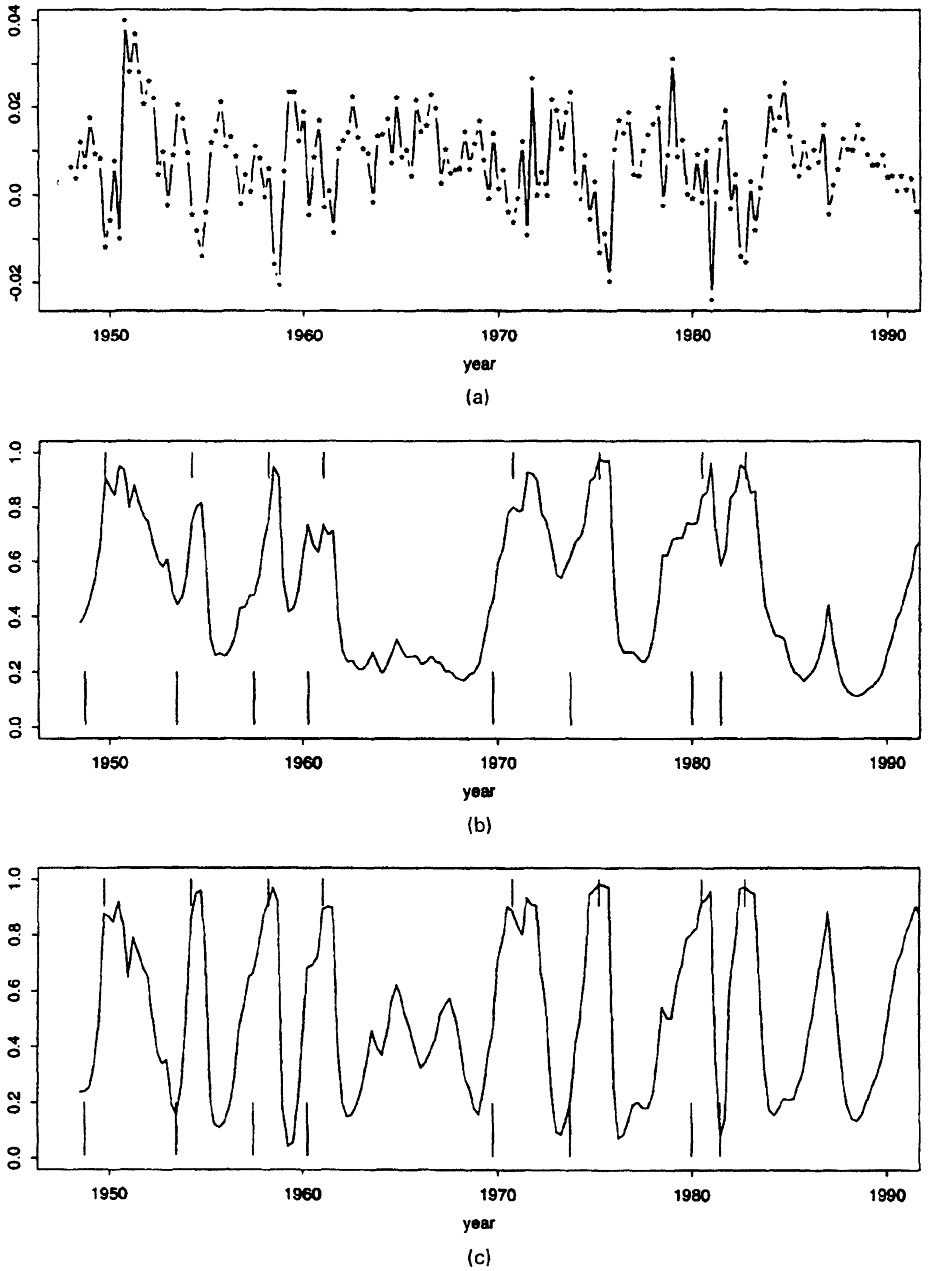

Fig 3. (a) Log(GNP growth); (b) posterior probability $\pi(I_i = 1|\mathcal{Y})$ using an AR(2) model with Markovian switching mechanism; (c) posterior probability $\pi(I_i = 1|\mathcal{Y})$ using an AR(2) model with balanced switching mechanism

and the date of a peak is usually later than the time that the posterior probability of expansion hits its local peak.

ACKNOWLEDGEMENTS

Rong Chen's research was supported in part by National Science Foundation (NSF) grant DMS-9301193 and Jun S. Liu's research was supported in part by NSF grant DMS-9404344. Part of the manuscript was prepared when Jun S. Liu was Assistant Professor, Department of Statistics, Harvard University. The authors thank Professor Ruey Tsay for providing the GNP data, Professor David Ellis for providing the phases of the business cycle and the referees for helpful suggestions.

REFERENCES

Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis.* Reading: Addison-Wesley.

Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992) Hierarchical Bayesian analysis of changepoint problems. *Appl. Statist.*, **41**, 389–405.

Chen, R. and Tsay, R. S. (1993) Functional-coefficient autoregressive models. *J. Am. Statist. Ass.*, **88**, 298–308.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

De Veaux, R. D. (1989) Mixtures of linear regressions. *Comput. Statist. Data Anal.*, **8**, 227–245.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc.* B, **56**, 363–375.

Escobar, M. D. (1994) Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Ass.*, **89**, 268–277.

Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions.* London: Chapman and Hall.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

Geweke, J. and Terui, N. (1991) Bayesian threshold autoregressive models for nonlinear time series. *Working Paper 483.* Research Department, Federal Reserve Bank of Minneapolis, Minneapolis.

Goldfeld, S. M. and Quandt, R. E. (1973) A Markov model for switching regression. *J. Econometr.*, **1**, 3–16.

Good, I. J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Hamilton, J. D. (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hosmer, Jr, D. W. (1974) Maximum-likelihood estimates of the parameters of a mixture of two regression lines. *Communs Statist.*, **3**, 137–148.

Irwin, M., Cox, N. and Kong, A. (1994) Sequential imputation for multilocus linkage analysis. *Technical Report.* Department of Statistics, University of Chicago, Chicago.

Kong, A., Liu, J. S. and Wong, W. H. (1994) Sequential imputations and Bayesian missing data problems. *J. Am. Statist. Ass.*, **89**, 278–288.

Lavine, M. and West, M. (1992) A Bayesian method for classification and discrimination. *Can. J. Statist.*, **20**, 451–461.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. and Wootton, J. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Liu, J. S. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Statist. Ass.*, **89**, 958–966.

——(1995) Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.*, to be published.

Lo, A. Y. (1987) A large sample study of the Bayesian bootstrap. *Ann. Statist.*, **15**, 360–375.

MacEachern, S. M. (1994) Estimating normal means with a conjugate style Dirichlet process prior. *Communs Statist. Simuln Computn*, **23**, 727–741.

McCulloch, R. E. and Tsay, R. S. (1992) Statistical inference of macro-economic time series via Markov switching models. *Technical Report*. Graduate School of Business, University of Chicago, Chicago.

Morris, C. N. (1983) Parametric empirical Bayes inference: theory and applications (with discussion). *J. Am. Statist. Ass.*, **78**, 47–65.

Quandt, R. E. and Remsey, J. B. (1978) Estimating mixtures of normal distributions and switching regressions. *J. Am. Statist. Ass.*, **73**, 730–738.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Render, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Robbins, H. (1955) An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 157–163. Berkeley: University of California Press.

Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.

——(1987) *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Shumway, R. H. and Stoffer, D. S. (1991) Dynamic linear models with switching. *J. Am. Statist. Ass.*, **86**, 763–769.

Smith, A. F. M. and Makov, U. E. (1978) A quasi-Bayes sequential procedure for mixtures. *J. R. Statist. Soc.* B, **40**, 106–112.

Tanner, M. (1993) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd edn. New York: Springer.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528–550.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Tong, H. (1983) Threshold models in nonlinear time series analysis. *Lect. Notes Statist.*, **21**.

West, M. (1992) Modelling with mixtures (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 503–524. Oxford: Oxford University Press.