Bayesian Model Monitoring

By MIKE WEST†

University of Warwick, UK

[Received February 1985. Final revision June 1985]

SUMMARY

A simple method of monitoring the predictive performance of a class of Bayesian models is introduced. The models involve sequential analyses of sequences of observations and are appropriate for a variety of monitoring and forecasting applications. For any given model, the monitoring technique is based on a comparison of the predictive ability of the model, measured by the observed values of predictive densities, with that of a single alternative. The alternative is constructed sequentially and is designed, for the class of models considered, as a relatively general yet neutral alternative to the original. The monitor is used as a general diagnostic tool to detect and assess discrepancies between the data and predictions generated from the model; particular sources of such model failure of interest are the occurrence of outliers and structural change in the series. An illustration is provided and a simple method of automatically coping with outliers and adapting to change is outlined.

Keywords: SEQUENTIAL BAYESIAN MODELLING; FORECASTING; MODEL ASSESSMENT; PREDICTIVE ABILITY; BAYES’ FACTORS; CUSUM; CHANGE POINT; OUTLIER REJECTION

1. INTRODUCTION

Bayesian modelling often naturally involves a sequential processing of observations. The general notion of progressively revising inferences and predictive statements as new information and observations are obtained is fundamental to the Bayesian approach. A sequential analysis also offers a modeller considerable flexibility in assessing individual observations and in continuously monitoring the predictive performance of a given model. In this paper, a new approach to sequential model monitoring is introduced. For a wide class of models, the aim is to provide an easily implemented diagnostic tool to detect and assess model/data discrepancies that may arise from a range of possible model failure modes.

The basic framework assumes a given, nominal model, referred to as the standard model, whose predictive performance, or fit, is to be assessed. The emphasis in sequential analyses is often on current and future behaviour of the observation series. Accordingly, the model monitor is specifically designed to be sensitive to local model failure, as is desirable in, for example, quality and inventory control, and time series forecasting. In these, and other application areas, one primary source of lack of fit, or standard model failure, is that due to structural changes in the series of observations. Such changes may be gradual or abrupt in nature and lead to sustained departures of the data from predictions in the standard model. Other possible sources of model failure, such as outliers or rare observations and extra stochastic variation, are also of importance and are discussed further below.

The Bayesian model monitor developed here is based on comparisons of predictions from the standard model with those from a single alternative model. The latter is constructed sequentially, observation by observation, and is designed as a relatively general and neutral alternative to the specific standard. In essence, the alternative is similar in form to the standard but allows for changes in the values of parameters characterizing the latter. The changes
allowed for are designed to be consistent with the types of structural changes expected in the data, yet the alternative model retains neutrality as to the direction and magnitude of such changes. As a by-product, the monitor is also effective in identifying outliers and other causes of model failure. Extra stochastic variation in the data, for example, may be ascribed to a sequence of changes or drifting in parameter values.

The approach here is Bayesian. Model fit is assessed using predictive distributions and the constructed alternative models have no non-Bayesian analogues in general. However, the resulting monitoring techniques have much in common with standard control schemes currently in wide use—in particular, sequential probability ratio tests (Page, 1954) and ‘backward cusums’ (Harrison and Davies, 1964). In common with such techniques, attention is focused to the ‘testing’ of a standard model. Until a monitor indicates model failure, the standard model is viewed as satisfactory. Questions about what should be done on model failure are touched on in Section 5 in a simple and special class of models. The use of more general elaborations of a standard model that provide explanations of the model failure are not considered here. Such elaborations for a wide class of dynamic forecasting models are developed in a related paper (West and Harrison, 1986).

2. SEQUENTIAL MODEL ASSESSMENT

2.1. Predictive Ability and Bayes’ Factors

Attention is restricted to univariate sequences of observations for which the sampling model at each point depends essentially on a single scalar parameter. The observations are denoted by \( \{ Y_t; t = 1, 2, \ldots \} \) where the index \( t \) will often, though not always, refer to time, and the realized value of \( Y_t \) is \( y_t \). At time \( t \), the history of the series is \( D_{t-1} \); thus \( D_t = \{ D_{t-1}, y_t \} \). Actually, \( D_{t-1} \) is viewed as sufficiently summarizing all the information available to and deemed relevant by the modeller, including, for example, known values of regressor variables, etc. The dependency of \( D_{t-1} \) on such information is suppressed for notational convenience. The standard model is defined sequentially. At time \( t \), the standard model for \( Y_t \) given \( D_{t-1} \) is denoted by \( S_t \) and involves: (a) a sampling distribution with density \( p(y_t | \eta_t) \) depending on the unknown, continuous parameter \( \eta_t \) (and possibly \( D_{t-1} \)); (b) a prior distribution for \( \eta_t \) given \( D_{t-1} \) with density \( p(\eta_t | D_{t-1}) \). The (prior or marginal) predictive distribution for \( Y_t \) given \( D_{t-1} \) then has density

\[
p(Y_t | D_{t-1}) = \int p(Y_t | \eta_t)p(\eta_t | D_{t-1})d\eta_t, \tag{2.1}
\]

the integral being over the parameter space for \( \eta_t \), a subset of \( \mathbb{R} \). The observed value \( p(y_t | D_{t-1}) \) is the fundamental measure of predictive ability for the standard model—the model likelihood.

The central practical problem in model assessment is that of constructing suitable alternatives to compare with \( S_t \); \( p(y_t | D_{t-1}) \) may be small—how small? Suppose that, at time \( t \), a suitable single alternative \( A_t \) exists having a predictive density \( p_A(Y_t | D_{t-1}) \). With reference to \( A_t \), the predictive fit of \( S_t \) is gauged using the model likelihood ratio, or Bayes’ factor

\[
H_t = \frac{p(y_t | D_{t-1})}{p_A(y_t | D_{t-1}).} \tag{2.2}
\]

Small values \( (< 1) \) of \( H_t \) indicate poor predictive performance of \( S_t \) and, if \( A_t \) is accepted as a sufficiently plausible alternative, \( S_t \) is discredited.

For the sequence of observations \( Y_1, Y_2, \ldots, Y_t \), overall model likelihoods may be calculated sequentially; \( S_1, S_2, \ldots, S_t \), have model likelihood given by the joint predictive density

\[
p(y_1, \ldots, y_t | D_0) = p(y_1 | D_{t-1})p(y_1, \ldots, y_{t-1} | D_0). \tag{2.3}
\]

A similar expression holds for the sequences of alternative models \( A_1, \ldots, A_t \), leading to an overall Bayes’ factor

\[
W_t = \frac{p(y_1, \ldots, y_t | D_0)}{p_A(y_1, \ldots, y_t | D_0)} = \frac{p(y_1 | D_{t-1})p(y_1, \ldots, y_{t-1} | D_0)}{p_A(y_1 | D_{t-1})p_A(y_1, \ldots, y_{t-1} | D_0)} = H_t W_{t-1}. \tag{2.3}
\]
$W_t$ is the basic tool in model assessment. In the practical monitoring context, however, where the focus is on local model fit, the global measure $W_t$ alone is not sufficient; other diagnostic aids are needed.

2.2. Cumulative Bayes’ Factors

Suppose that, at time $t$, $W_{t-1} > 1$ so that the standard model is viewed as having been satisfactory to date, but that $y_t$ is extremely discrepant with $H_t$ is very small. Then, even though $y_t$ is obviously discrepant, the weight of evidence $W_{t-1}$ in favour of the standard model so far may be so large that $W_t = H_t W_{t-1}$ may be greater than one, even considerably greater. In this case, the evidence against $S_t$ is masked by that in favour of $S_t$, ..., $S_{t-1}$; the local model failure is not evidenced by $W_t$ but by $H_t$. In the case of change, where $H_t$, $H_{t+1}$, ... are all small, the evidence of standard model failure may take several sampling intervals to become apparent in the Bayes’ factor. Furthermore, in cases of small or gradual change, the individual Bayes’ factors $H_t$, $H_{t+1}$, ... will be small but they may not be small enough to signal model failure individually. Accordingly, to focus on possible local model failure, it is necessary to consider not only the sequences $W_t$ and $H_t$, but also the evidence for the standard model for groups of the most recent consecutive observations.

Generally, define the cumulative Bayes’ factor

$$W_{t}(k) = H_t H_{t-1} \cdots H_{t-k+1}, \quad (k = 1, 2, \ldots, t).$$

Clearly $W_t(t) = W_t$, $W_t(1) = H_t$, and, for $t > 1$,

$$W_{t}(k) = H_t W_{t-1}(k - 1).$$

$W_t(k)$ assesses the fit of the most recent $k$ observations. The possibility that changes may have occurred at some time before $t$ may be investigated using these factors. A single small $H_t = W_t(1)$ provides a warning of a potential outlier or the onset of change; a small $W_t(k)$ for $k > 1$ suggests a possible change in the past. To focus on the most likely point of possible change, it is necessary to identify the most discrepant group of recent, consecutive observations. This involves calculating, at time $t$, the quantity

$$V_t = \min_{1 \leq k \leq t} W_t(k).$$

For each $t$, there exists $l_t$ such that $V_t = W_t(l_t)$. Clearly, $y_t$, $y_{t-1}$, ..., $y_{t-l_t+1}$ is the most discrepant group of observations of interest. $V_t$ can be efficiently calculated using

$$V_t = \min \{ W_t(1), \min_{2 \leq k \leq t} W_t(k) \} = H_t \min \left\{ 1, \min_{1 \leq j \leq t-1} W_{t-1}(j) \right\},$$

leading to the neat recursion

$$V_t = H_t \min \{ 1, V_{t-1} \}. \quad (2.4)$$

The related run-length parameter $l_t$ may also be calculated sequentially from

$$l_t = \begin{cases} l_{t-1} + 1, & \text{if } V_{t-1} < 1; \\ 1, & \text{if } V_{t-1} \geq 1. \end{cases}$$

The focus on local behaviour of the series is evident in $V_t$. If, at time $t$, the evidence favours the standard description of the series so that $V_{t-1} \geq 1$, then $V_t = H_t$, and decisions about possible inadequacies are based on $y_t$ alone. If $H_t$ is very small, then $y_t$ is a potential outlier or may indicate the onset of change. If, however, the evidence is against the standard models before time $t$ so that $V_{t-1} < 1$, then the cumulative Bayes’ factor is further decreased, and the run-length $l_t$ increased by one, if $H_t < 1$. If $H_t > 1$, then the cumulative factor is increased, possibly even to exceed one. Note that, starting from a factor of one, a relatively slow or
gradual change will be detected as consecutive, fairly small values of the $H_t$ sequence are cumulated. Furthermore, the run-length $l_t$ provides an indication of the most likely point of onset of change.

Finally, it is of interest to note that $V_t$ defines what may be termed a ‘Bayesian cusum’. If $Z_t = \ln(V_t)$, then

$$Z_t = \ln(H_t) + \min\{0, Z_{t-1}\}.$$  

$Z_t$ has the form of a standard cusum test statistic based on a sequential probability test. The operating characteristics of Bayesian monitoring schemes (such as expected run length for specific, plausible model failures) may be investigated using the methods of Kemp (1971) although this is not done here.

3. A GENERAL, NEUTRAL ALTERNATIVE

In constructing $A_t$, it is not necessary to consider specific departures from the standard model that are potential causes of an observed discrepancy between $y_t$ and $S_t$. All that is required for a suitably neutral alternative predictive density $p_A(Y_t | D_{t-1})$ is that it has similar location and shape characteristics to the standard $p(Y_t | D_{t-1})$, whilst being more diffuse. Such a density may be constructed directly, in a fashion similar to the standard, via

$$p_A(Y_t | D_{t-1}) = \int p(Y_t | \eta_t)p_A(\eta_t | D_{t-1})d\eta_t,$$

where the alternative prior density $p_A(\eta_t | D_{t-1})$ has the same functional form as the standard prior, similar location but greater spread.

A relatively general prescription for $p_A(\eta_t | D_{t-1})$, based on the use of power discounting as in Smith (1979), is adopted. Basically, for any $\delta_t$, with $0 < \delta_t < 1$, the alternative prior

$$p_A(\eta_t | D_{t-1}) \propto p(\eta_t | D_{t-1})^{\delta_t}$$  

(3.1)

has location and shape similar to the standard prior but is more diffuse; the factor $\delta_t$ effectively discounts the information summarized by the standard prior to produce an alternative having a smoother, flatter form. Before developing this power discount method, the dependence on parametrization is noted. As an example, suppose that $Y_t$ is Bernoulli with probability $\pi_t$ and $\eta_t = \ln\{\pi_t/\{1-\pi_t\}\}$. The standard prior for $(\pi_t | D_{t-1})$ is the Beta form $B[a, b - a], (b > a)$, so that $\ln p(\eta_t | D_t) = c + a\eta_t - b\ln[1 + \exp(\eta_t)]$. For a given discount $\delta_t$, the power transformation (3.1) may be applied to either $\eta_t$ or $\pi_t$. In the first case, $p_A(Y_t | D_{t-1})$ has the same form as the standard prior for $\eta_t$ but with $[a, b]$ replaced by $[a\delta_t, b\delta_t]$; the corresponding prior for $\pi_t$ is $B[a\delta_t, (b - a)\delta_t]$. In the second case, the power transformation applied to $p(\pi_t | D_{t-1})$ leads to an alternative in which $[a, b]$ is replaced by $[1 + (a - 1)\delta_t, 2 + (b - 2)\delta_t]$, so that here and, of course, more generally, the power discounting approach is parametrization dependent. Typically, this is not a practical problem but in this example it is. Note that $p(Y_t = 1 | D_{t-1}) = E[\pi_t | D_{t-1}] = a/b$, so that the effect of using power discounting on the prior for $\eta_t$ as described above is to lead to $p_A(Y_t = 1 | D_{t-1}) = a\delta_t/\{b\delta_t\} = p(Y_t = 1 | D_{t-1})$, and $H_t = 1$ for all $t$. Discounting $\pi_t$, however, gives $p_A(Y_t = 1 | D_{t-1}) = [1 + \delta_t(a - 1)]/[2 + \delta_t(b - 2)]$ and this leads to a suitable Bayes’ factor to assess $S_t$.

4. EXPONENTIAL FAMILY MODELS

4.1. General Structure and Special Cases

A typical distribution in the exponential family has a density of the form

$$p(Y_t | \eta_t) = h(Y_t, \phi_t)\exp\{\phi_t[Y_t\eta_t - a(\eta_t)]\},$$  

(4.1)

where $a(.)$ and $h(.)$ are known functions, $\phi_t$ is a known scale, or precision parameter and $\eta_t$, the natural parameter, defines location via $\mu_t = E[Y_t | \eta_{t-1}] = a(\eta_t)$. The classes of Bayesian
models referred to below all utilize conjugate prior distributions for \((\eta_t | D_{t-1})\), and attention here is restricted to such models. A typical prior has a density of the form

\[ p(\eta_t | D_{t-1}) = k(a_t, b_t) \exp \{ a_t \eta_t - b_t a(\eta_t) \}. \]  \hspace{1cm} (4.2)

The exponent here may be written as \( b_t [m_t \eta_t - a(\eta_t)] \) where \( m_t = a_t/b_t \) defines the location of the prior. The prior mode of \( \eta_t \) satisfies \( m_t = \tilde{a}(\eta_t) \) and, typically, \( E[\mu_t | D_{t-1}] = m_t \). The precision parameter \( b_t \) provides a contrast of the strength of information in the likelihood (3.1) for \( \eta_t \) with that summarized by the prior; this is measured by the relative precisions, or signal : noise ratio \( b_t/\phi_t \).

**Example 4.1**

Simple static random sampling models involve \( \eta_t = n \) for all \( t \). With a prior \( p(\eta | D_0) \) having parameters \([a_0, b_0]\), then, for \( t > 0 \) \([a_t, b_t]\) are calculated sequentially using \( a_t = a_{t-1} + \phi_{t-1} y_{t-1} \) and \( b_t = b_{t-1} + \phi_{t-1} \).

**Example 4.2**

West, Harrison and Migon (1985) discuss the class of dynamic generalized linear models (D.G.L.M.’s) and develop a novel approach to modelling appropriate for a variety of time series, forecasting and dynamic, non-linear regression applications. These models include as special cases static linear and generalized models, dynamic linear normal models, and simple steady models similar to those of Smith (1979). In a typical D.G.L.M. a monotone transform of \( \eta_t, g(\eta_t) \) say, is related to a linear function of a known regressors \( \beta_1 \theta_t \), where \( \theta_t \) is a time-varying ‘regression’ vector. For such a model, the analysis at time requires the specification prior mean and covariance matrix of \( \theta_t \), denoted by \((\theta_t | D_{t-1}) \sim [\alpha_t, \beta_t]\). These moments are used to define the conjugate prior by a mapping onto the parameters \([a_t, b_t]\) of the prior for \( \eta_t \). Sequential learning about \( \theta_t \) involves updating to the posterior \( p(\eta_t | D_t) \) and then ‘feeding back’ the information provided by \( y_t \) to a posterior for \((\theta_t | D_t)\) using a system of sequential updating equations similar to those in dynamic linear normal models.

In the first example, \( \eta_t \) is supposedly constant in time. A monitoring scheme suited to detecting deviation from this hypothesis might be used, for example, in quality control and inventory applications. The models in Example 4.2 provide for smooth changes over time in \( \theta_t \). Poor predictions from the model indicate the possibility of more marked, abrupt changes having occurred. In all models, however, a lack of fit may be due to other, unspecified features such as extra random variation or outliers. An outlier will be detected by a monitor but, based on information available at time \( t - 1 \) alone, an outlying \( y_t \) cannot be distinguished from the onset of change at time \( t \). Outlier handling is discussed further in Section 5 below.

4.2. **Predictive Distributions**

In standard model \( S_t \), the predictive density for \((Y_t | D_{t-1})\) is defined via (4.1) and (4.2) as

\[ p(Y_t | D_{t-1}) = h(Y_t, \phi_t)k(a_t, b_t)/k(a_t + \phi_t Y_t, b_t + \phi_t). \]  \hspace{1cm} (4.3)

The location of this density depends primarily on that of the prior (4.2); typically \( E[Y_t | D_{t-1}] = E[\mu_t | D_{t-1}] = m_t \). The variance of the predictive distribution may be shown, generally, to be given by \( V[Y_t | D_{t-1}] = \sigma_t E[\tilde{a}(\eta_t) | D_{t-1}] \), where \( \sigma_t = \{1/\phi_t + 1/b_t\} \). The expectation term here is with respect to the prior (3.2) and so is a function of both \( a_t \) and \( b_t \).

Typically, however, the first term \( \sigma_t \) involves the primary dependence on the prior scale parameter \( b_t \), and provides a general measure of spread referred to here as the *scale-sum* of the predictive density.

Using the power discounting (3.1) for the prior (4.2), the alternative prior satisfies \( \ln p_A(\eta_t | D_{t-1}) = c + d_t[m_t \eta_t - a(\eta_t)] \), where \( d_t = \delta t b_t \) for some discount factor \( \delta_t \) \((0 < \delta > 1)\). The alternative predictive distribution then has a density of the same form as (4.3) with the \( b_t \) replaced by \( d_t \). Note that the locations of the two predictive densities are similar, with
The scale-sum of the alternative model is simply $1/\phi_t + 1/(\delta_t b_t)$ which is clearly greater than $\sigma_t$ since $\delta_t < 1$.

The discount factor $\delta_t$ should be chosen to ensure that, in general terms, the relative spreads of the standard and alternative predictive distributions be constant in time. In particular, a requirement of constant relative scale-sum values implies that, for some fixed $\rho, (0 < \rho < 1)$, $1/\phi_t + 1/b_t = \rho [1/\phi_t + 1/(\delta_t b_t)],$ which leads to $\delta_t = \rho/[1 + (1 - \rho)(b_t/\phi_t)].$ Since $0 < \rho < 1$, it is clear that $\delta_t < \rho$ for all $t$. When $b_t/\phi_t$ is large, so that the standard prior is highly concentrated relative to the likelihood, then $\delta_t$ is much less than $\rho$. Thus a high degree of discounting of the prior is used to construct an alternative predictive density that is appreciably more diffuse than the standard. In the limit as $1/b_t$ tends to zero indicating a degenerate standard prior, $\delta_t$ also tends to zero but $d_t = \delta_t b_t$ tends to $\rho \phi_t/(1 - \rho).$ Hence the alternative model is based on a non-degenerate prior. Conversely, as the ratio $b_t/\phi_t$ tends to zero, so that the prior is much more diffuse than the likelihood, then $\delta_t$ tends to the constant discount $\rho$.

5. SIMPLE MODEL MONITORING AND ADAPTATION

5.1. Signalling Model Failure

In practice the cumulative Bayes’ factor $V_t$ of (2.4) is used as a (non-standard) sequential probability ratio test statistic. This simple mode of operation involves monitoring the standard model until the evidence of inadequacy, or model failure, as measured by a small value of $V_t$, is sufficiently great to warrant intervention. A threshold value $\tau$ for $V_t$ is specified; the standard model is considered satisfactory until such a time as $V_t < \tau$. Notice that $\tau$ is not assumed to depend on either the time index $t$ or the run-length parameter $l_t$. In this the Bayesian monitor differs from standard sampling theoretic schemes (Page, 1954) in which $\tau$ would be chosen as a lower percentile of the sampling distribution of $V_t$ under the ‘null hypothesis’ of the standard models. Typically this distribution will depend on both $t$ and $l_t$. Operational characteristics of the Bayesian monitor in any particular application now depend on the two constant parameters $\rho$ and $\tau$. Some guide as to the choice of values are obtained by consideration of the following special case.

Example 5.1

Suppose that $(Y_t | \eta_t) \sim N[\eta_t, 1/\phi_t]$, $(\eta_t | D_{t-1}) \sim N[m_t, 1/b_t]$, so that, in $S_n$, $(Y_t | D_{t-1}) \sim N[m_t, \sigma_t]$, whereas, in $A_t$, $(Y_t | D_{t-1}) \sim N[m_t, \gamma_t]$ where $\gamma_t = 1/\phi_t + 1/(b_t \delta_t)$. Using the value of $\delta_t$ given in (4.4) leads to $\gamma_t = \rho \gamma_t$, and, in this case, $H_t = \exp[-0.5(1-\rho)e_t^2]/\sqrt{\rho}$, where $e_t$ is the standardized prediction residual from $S_t$, $e_t = (y_t - m_t)/\sqrt{\sigma_t}$. Now, the predictive standard deviation in $A_t$ is a factor of $1/\sqrt{\rho}$ times that in $S_t$. An increase in standard deviation of 2.5 times leads to $\rho \approx 0.15$ whereas one of 3 times gives $\rho \approx 0.1$. Suitable values of $\rho$ lie in the range (0.1, 0.3); the monitor is fairly robust to the particular value within this range. Some guidance may be obtained by considering the cross-over point of the predictive densities. This is the point at which $H_t = 1$ so that the single observation $y_t$ provides no evidence to distinguish between $S_t$ and $A_t$. To be ambivalent between the two models when $|e_t| = 1.5$ in the standard model, then $H_t = 1$ implies that $-\ln(\rho) = (1.5)^2(1 - \rho);$ this leads to $\rho \approx 0.15$. This rough value is recommended for routine use. A larger value will lead to increased tolerance of $S_t$ with residuals greater than 1.5 in absolute value being ‘acceptable’; a smaller value clearly has the reverse effect.

Consideration of the single observation $y_t$ also guides the choice of the threshold parameter $\tau$ of the monitor. At the threshold point for a single observation $V_t = H_t = \tau$ and $y_t$ alone provides sufficient evidence to ‘reject’ $S_t$ as inadequate. With $|e_t| = k$, then $H_t = \tau = \exp[-0.5(1-\rho)k^2]/\sqrt{\rho}$. Thus, if $\rho = 0.15$, then $k = 2$ implies that $\tau \approx 0.47$ whereas $k = 2.5$ gives $\tau \approx 0.18$. Typically, a value of $\tau \approx 0.4$ is recommended for routine use. This is consistent
with the guidelines of Jeffreys (1961, Appendix B) for interpreting the size of observed Bayes’ factors; a value of $10^{-1/2}$ or less indicates evidence against the standard model.

**Example 5.2**

As a simple, non-normal illustration, suppose that the standard model has $Y_t = X_t/\phi_t$ where $(X_t|\eta_t) \sim \text{Bin}[\phi_t, \pi_t]$, with $\eta_t = \ln[\pi_t/(1 - \pi_t)]$. The conjugate prior for $\pi_t$ is Beta, $(\pi_t|D_{t-1}) \sim B[a_t, b_t - a_t]$, with precision parameter $b_t (b_t > a_t)$, mean $m_t = a_t/b_t$, and variance $m_t(1 - m_t)/(1 + b_t)$. Smith (1980) discusses a general, non-sequential approach to change-point problems and examines a binomial model for the sequence of observations in Table 1.

### Table 1

**Binomial data**

<table>
<thead>
<tr>
<th>$t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_t$</td>
<td>21</td>
<td>36</td>
<td>44</td>
<td>30</td>
<td>52</td>
<td>45</td>
<td>48</td>
<td>57</td>
<td>48</td>
<td>22</td>
<td>20</td>
<td>21</td>
<td>20</td>
</tr>
<tr>
<td>$X_t$</td>
<td>12</td>
<td>26</td>
<td>31</td>
<td>24</td>
<td>28</td>
<td>34</td>
<td>39</td>
<td>46</td>
<td>41</td>
<td>19</td>
<td>17</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>$Y_t$</td>
<td>0.57</td>
<td>0.72</td>
<td>0.71</td>
<td>0.80</td>
<td>0.54</td>
<td>0.76</td>
<td>0.81</td>
<td>0.85</td>
<td>0.86</td>
<td>0.85</td>
<td>0.81</td>
<td>0.80</td>
<td></td>
</tr>
</tbody>
</table>

The problem is one of identifying change points in the series as modelled by changes in $\pi_t$. The basic conclusion of Smith is that the evidence strongly suggests a change at $t = 5$ followed by one at $t = 6$. A simple, sequential alternative to the elaborate model of Smith is to assume a common binomial distribution for the observations, $\pi_t = \pi$ for all $t$, and to monitor for evidence of model failure. This approach was taken starting with a uniform prior for $\pi$, $(\pi_t|D_0) \approx B[1, 1]$, the discount $\rho = 0.15$, and the threshold $\tau = 0.4$. The alternative model $A_t$ was based on power discounting for $\pi_t$. The values of the resulting Bayes’ factors and run-length are given in Table 2. Also given are two features of the posterior distributions for $(\pi|D_t)$, for each $t$: namely the sequence of posterior precision parameters $b_{t+1} = b_t + \phi_t$, and posterior means $m_{t+1} = (a_t + X_t)/b_{t+1}$.

### Table 2

**Monitoring the binomial data**

<table>
<thead>
<tr>
<th>$t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_t$</td>
<td>4.00</td>
<td>1.42</td>
<td>2.45</td>
<td>1.20</td>
<td>0.39</td>
<td>1.28</td>
<td>0.51</td>
<td>0.75</td>
<td>0.37</td>
<td>1.00</td>
<td>1.40</td>
<td>2.16</td>
<td>2.37</td>
</tr>
<tr>
<td>$V_t$</td>
<td>4.00</td>
<td>1.42</td>
<td>2.45</td>
<td>1.20</td>
<td>0.39</td>
<td>0.51</td>
<td>0.26</td>
<td>0.19</td>
<td>0.07</td>
<td>0.07</td>
<td>0.10</td>
<td>0.21</td>
<td>0.51</td>
</tr>
<tr>
<td>$l_t$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>$m_{t+1}$</td>
<td>0.56</td>
<td>0.66</td>
<td>0.70</td>
<td>0.71</td>
<td>0.66</td>
<td>0.68</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
<td>0.74</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>$b_{t+1}$</td>
<td>23</td>
<td>59</td>
<td>103</td>
<td>133</td>
<td>185</td>
<td>230</td>
<td>278</td>
<td>335</td>
<td>383</td>
<td>405</td>
<td>425</td>
<td>446</td>
<td>466</td>
</tr>
</tbody>
</table>

From the $H_t$ sequence there is clear evidence of model failure at $t = 5$. Thereafter $l_t$ continually increases indicating poor subsequent predictive performance of the sequence of standard models. Notice that, although $V_t$ decreases after $t = 6$, there is an eventual increase as the posteriors for $\pi$ are updated and slowly adapt to the later observations. This adaptation is shown in the posterior means $m_{t+1}$; it is, however, extremely slow due to the build up of information about $\pi$ from earlier observations evident in the precisions $b_{t+1}$.

5.2. **Simple Outlier Rejection and Model Adaptation**

When changes in parameter values are the primary cause of standard model failure, an automatic method of *adapting* to a change is often required. If the monitor signals a change, the standard model should be allowed to adapt as soon as possible in order to improve further predictions. In Example 5.2 above, the model failure at $t = 5$ is detected but the standard model is unchanged and subsequent poor predictions are evident. Intervention is required in order to discount the information about $\pi$, providing, for the next time point, a more diffuse
prior so that the posterior will depend more heavily on the next observation. If the possible changes of interest at time \( t \) are modelled by changes in \( \eta \), this loss of information would be naturally and simply achieved by replacing the standard prior with the more diffuse prior alternative. This requires, however, prior knowledge of the point of change. Furthermore, it would be potentially disastrous if a discrepant observation turned out to be an outlier rather than a change point.

Consider the possibility that a signal of model failure is due to an outlier rather than a change. In Example 5.2, for instance, the observation at time \( t = 5 \) may be considered an outlier. More generally, \( y_t \) is a potential outlier if the monitor signals standard model failure based on the single observation \( l_t = 1 \), rather than a group \( l_t > 1 \). If a single discrepant point is an outlier and should be ignored) (or rejected), then the above suggestion of automatically switching to the alternative prior requires modification since the corresponding posterior will then depend heavily on the erroneous information provided by the outlier. Of course a single discrepant observation cannot be judged without further information; it may be an outlier, but it may indicate the onset of change. In the first case, it should be rejected. In the second, rejection of the observation implies losing creditable information but further observations will confirm the change and provide information about the direction and magnitude. Accordingly, the following, automatic method of detecting and rejecting outliers and adapting to change in simple models is proposed.

At time \( t \), whatever has occurred previously, proceed as follows:

(A) (Assessment of \( S_t \)). Calculate the single Bayes' factor \( H_t \). If \( H_t \geq \tau \), then \( y_t \) is viewed as consistent with \( S_t \); proceed to (B) to assess the possibilities of model failure before time \( t \). If, on the other hand, \( H_t < \tau \), then \( y_t \) is a potential outlier and should be rejected. However, the possibility of onset of a change must be allowed for too; proceed to (C) to achieve this.

(B) (Assessment of \( S_t, S_{t-1}, \ldots \)). Calculate the cumulative factor \( V_t \) and the corresponding run length \( l_t \) to assess the possibility of model failure before time \( t \). If \( V_t \geq \tau \), then the sequence of standard models is viewed as satisfactory; proceed to (D). Otherwise, \( V_t < \tau \) indicates change so proceed to (C). (n.b: A variant of this scheme that increases sensitivity to rather slow changes in the series is to signal failure and proceed to (C) whenever \( V_t < \tau \) or \( l_t > k \), for \( k \) about 3 or 4).

(C) (Allowing for change). Reject the single observation \( y_t \). To allow for change, set the posterior to \( p(\eta_t | D_t) = p_A(\eta_t | D_{t-1}) \) to increase the uncertainty about \( \eta_t \) and lead to greater adaptation to new data. Update the time index to \( t + 1 \) and return to (A).

(D) (Standard update). The sequence of standard models in satisfactory. Update as usual to the posterior \( p(\eta_t | D_{t-1}) \propto p(y_t | \eta_t) \), and thence to the prior for time \( t + 1 \), \( p(\eta_{t-1} | D_t) \). Update the time index from \( t \) to \( t + 1 \) and return to (A).

**Example 5.3**

Applied to the data in Example 5.2, the procedure leads to the rejection of \( y_2 \) as an outlier followed by adaptation to the change at \( t = 6 \). In this analysis \( l_t = 1 \) and \( V_t = H_t \) for each \( t \), indicating that the change point is adequately catered for. Table 3 provides the details, including the sequence of values of the discounts \( \delta_t \) generated. Recall that \( \delta_t < \rho = 0.15 \) and that \( \delta_t \) decreases as \( b_t \) increases.

**TABLE 3**

Outlier rejection and model adaptation for binomial data

<table>
<thead>
<tr>
<th>( \delta_t )</th>
<th>0.14</th>
<th>0.10</th>
<th>0.07</th>
<th>0.04</th>
<th>0.05</th>
<th>0.13</th>
<th>0.08</th>
<th>0.06</th>
<th>0.04</th>
<th>0.02</th>
<th>0.01</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_{y+t} )</td>
<td>0.56</td>
<td>0.60</td>
<td>0.71</td>
<td>0.74</td>
<td>0.78</td>
<td>0.80</td>
<td>0.80</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>( b_{y+t} )</td>
<td>23</td>
<td>59</td>
<td>103</td>
<td>133</td>
<td>7</td>
<td>52</td>
<td>100</td>
<td>157</td>
<td>205</td>
<td>227</td>
<td>247</td>
<td>268</td>
</tr>
<tr>
<td>( H_t )</td>
<td>4.00</td>
<td>1.42</td>
<td>2.45</td>
<td>1.20</td>
<td>0.39</td>
<td>3.23</td>
<td>2.08</td>
<td>2.49</td>
<td>1.62</td>
<td>2.09</td>
<td>2.47</td>
<td>2.90</td>
</tr>
</tbody>
</table>
The posteriors for $\pi$ adapt much more rapidly to the change as evidenced by the sequence of means. This is due to the intervention after detecting the outlier at time 5 that leads to a decrease in the precision $b_6$ from the original value of $133 + 52 = 185$ in Example 5.2 to $133(0.05) \approx 7$ here.

In general, if $y_t$ is rejected but further data indicates that it was, in fact, the start of a change, the procedure may be refined to then incorporate the information in $y_t$. Typically, however, this will have only a small effect and so is not considered here. The general procedure has been successfully applied in a variety of monitoring and forecasting applications; illustrations using dynamic Bayesian models appear in the related paper by West and Harrison (1986).

ACKNOWLEDGEMENTS

I am grateful to Jeff Harrison for discussion of the ideas in this paper, and to the referees for their comments on the original draft of the paper.

REFERENCES


