Mixture Models, Monte Carlo, Bayesian Updating and Dynamic Models

Mike West
Institute of Statistics & Decision Sciences
Duke University, Durham NC 27708, USA

Abstract

This paper reviews the development of discrete mixture distributions as approximations to priors and posteriors in Bayesian analysis, focusing on the development of simulation based techniques in sequential updating and the analysis of dynamic models. Adaptive density estimation techniques enable construction of mixtures of elliptical distributions useful as direct approximations and as importance sampling functions. Illustrations in sequential modelling are discussed.

1. Adaptive mixture modelling

1.1 Importance sampling and mixtures

With the recent mushrooming application of simulation based methods of numerical analysis, Bayesian analyses often involves the approximation of continuous prior and posterior distributions using discrete sets of points and associated weights based on a Monte Carlo approximation. West (1992a) introduced an adaptive importance sampling scheme to develop such discrete approximations, and methods to provide smooth reconstructions, in cases when the prior and likelihood functions separately, or the posterior directly, may be (at least approximately) evaluated up to irrelevant constants. Suppose \( p(\theta) \) is the continuous posterior density function for a continuous parameter vector \( \theta \). An approximating density \( g(\theta) \) is used as an importance sampling function (Geweke, 1989) as follows. Let \( \Theta = \{ \theta_j, \ j = 1,\ldots,n \} \) be a random sample from \( p(\theta) \), and define weights \( \Omega = \{ w_j, \ j = 1,\ldots,n \} \) by \( w_j = p(\theta_j)/(kg(\theta_j)) \), for each \( j \), where \( k = \sum_{j=1}^{n} p(\theta_j)/g(\theta_j) \). The weights are evaluated via \( w_j \propto p(\theta_j)/g(\theta_j) \), and then normalised to unit sum. Then inference under \( p(\theta) \) is approximated using the discrete distribution having masses \( w_j \) at \( \theta_j \), for each \( j = 1,\ldots,n \). The conditions on \( g(\theta) \) required to achieve reasonable approximations typically reduce to requiring \( g \) have the same support as \( p \) and that the tails of \( g \) be heavier than those of \( p \), so that variations on multivariate T distributions have become popular (Geweke, 1989). In West (1992a) mixtures of T distributions are proposed, one additional reason being that mixtures have the flexibility to represent the possibly quite complex and varied forms of (posterior) densities. This is done using kernel density estimation techniques. With an importance sampling function \( g(\theta) \) close to the true density \( p(\theta) \), kernel density estimation (or other smoothing techniques) provides continuous estimates of joint and marginal densities. In simple, univariate random sampling problems, kernel type density estimates have direct Bayesian interpretation as approximate predictive distributions in models based on mixtures of Dirichlet processes (West, 1990); multivariate analogues are similarly derivable (Erkanli, Müller and West, 1992). West (1992a) uses weighted variations on multivariate kernel estimates as importance sampling functions and, with some modification, to more directly estimate marginal densities of \( p(\theta) \). The basic idea is as follows.

Given a chosen importance sampling density \( g_0(\theta) \), the sample, of size \( n \), \( \Theta \) and associated weights \( \Omega \), the exact density \( p(\theta) \) may be approximated by a weighted kernel estimate of the form

\[
g_1(\theta) = \sum_{j=1}^{n} w_j d(\theta|\theta_j, \mathbf{V}h^2)
\]

where \( d(\theta|m, M) \) denotes a \( p \)-variate, elliptically symmetric density function (determining the ‘kernel’), with mode \( m \) and scale matrix \( M \), \( \mathbf{V} \) is an estimate of the variance matrix of \( p(\theta) \), (usually the Monte Carlo estimate based on \( \Theta \) and \( \Omega \)), and \( h \) is a ‘window-width’ smoothing parameter, depending on the Monte Carlo sample size \( n \). A key example has \( d(\theta|m, M) \) as the density of a multivariate T distribution with some \( a > 0 \) degrees of freedom, whence the density \( g_1(\theta) \) is a discrete mixture of \( n \) component T distributions. Conventional density estimation techniques (Silverman, 1986) choose the window width \( h \) as

Acknowledgement: Research reported here was partially supported by NSF under grants DMS-8903842 and DMS-9024793
a slowly decreasing function of \( n \), so that the kernel components are naturally more concentrated about the locations \( \theta_j \); for larger sample sizes, and the approach of West (1990) indicates how currently popular rules for choosing \( h \) may be interpreted within a Bayesian and, more importantly, model-based framework. If \( h \) does decay to zero at the right rate as \( n \) increases, then \( g_1(\theta) \) approaches \( p(\theta) \) as \( n \) increases. For moderate sample sizes, \( g_1(\theta) \) will tend to be overdispersed relative to \( p(\theta) \) if \( V \) is, as is usual, the Monte Carlo estimate of \( V[\theta] \) under \( p(\theta) \). This feature proves useful in developing kernel forms as importance density functions, though can be counter-balanced using ‘shrinkage’ of kernel locations to provide more direct approximation of the true density and its margins (West, 1992a).

Adaptive methods of posterior approximation based on the generalised kernel technique are now discussed. Note first, however, that there are obvious ways in which they may require modification in specific applications. A key issue concerns the possible patterns of local dependence exhibited by \( p(\theta) \), and the use in (1) of a global estimate of covariance. In many problems, this will be quite adequate, but it is not uncommon to encounter problems in which different regions of parameter space are associated with rather different patterns of dependence, so that a local estimate of covariance structure, with \( V \) in (1) varying with locale \( j \) and more heavily depending on \( \theta_j \) values near \( \theta_j \), is required. Other modifications involve the more formal modelling foundation for kernel type techniques that derive from the approach to density estimation using Dirichlet process mixtures of normal distributions (West, 1990; Escobar and West, 1992). Current developments of the dynamic modelling applications using these ideas are mooted in Müller (1993).

Finally, some of the modifications mentioned in West (1992a), especially a neat data augmentation trick used to effect dimensionality reduction and significantly reduce computations in multiparameter models, may be directly applied in sequential problems.

1.2 Adaptive importance sampling

Adaptive importance sampling describes any process by which the importance sampling distribution is sequentially revised based on information derived from successive Monte Carlo samples. Let \( g_0(\theta) \) be an initially chosen importance sampling function. For a sample size \( n \), this leads to points \( \Theta_0 = \{\theta_{0,j}, \ j = 1, \ldots, n_0\} \), and weights \( \Omega_0 = \{w_{0,j}, \ j = 1, \ldots, n_0\} \), and the summary

\[
\Gamma_0 = \{g_0, n_0, \Theta_0, \Omega_0\}. 
\]

Adaptive importance sampling suggests taking \( n_0 \) rather small, say several hundreds, and, based on the Monte Carlo outcome \( \Gamma_0 \), revising \( g_0(\theta) \) to a ‘better guess’. It is natural to use a mixture such as (1) as a second step importance sampling function, and the following adaptive routine arises:

1. Choose an initial importance sampling distribution with density \( g_0(\theta) \), draw a fairly small sample \( n_0 \) and compute weights, deducing the summary \( \Gamma_0 = \{g_0, n_0, \Theta_0, \Omega_0\} \). Compute the Monte Carlo estimates \( \hat{\theta}_0 \) and \( \hat{\Omega}_0 \) of the mean and variance of \( p(\theta) \).
2. Construct a revised importance function \( g_1(\theta) \) using (1) with sample size \( n_0 \), points \( \theta_{0,j} \), weights \( w_{0,j} \), and variance matrix \( \hat{\Omega}_0 \).
3. Draw a larger sample of size \( n_1 \) from \( g_1(\theta) \), and replace \( \Gamma_0 \) with \( \Gamma_1 = \{g_1, n_1, \Theta_1, \Omega_1\} \).
4. Either stop, and base inferences on \( \Gamma_1 \), or proceed, if desired, to a further revised version \( g_2(\theta) \), constructed similarly. Continue as desired.

Even though the initial \( g_0(\theta) \) may poorly represent \( p(\theta) \), successive refinement through smaller samples can, and usually does, mean that, after one or two revisions, the resulting kernel estimate is a much better approximation to \( p(\theta) \). Hence, once the process of refinement is terminated, a much smaller sample size is necessary for the desired accuracy of approximation. Often just one refinement is sufficient to adjust a very crude approximation, \( g_0(\theta) \), say a single multivariate T density, to a mixture \( g_1(\theta) \) of, say, 500 T densities, that much more closely represents the true \( p(\theta) \). In approximating moments and probabilities, a Monte Carlo sample of two or three thousand draws from \( g_1(\theta) \) may do as well as, or better than, a sample of two or three times that from the original \( g_0(\theta) \). Useful diagnostic information is generated in this process at each stage, such as the configuration of points in each dimension of the parameter space, and the distributions of weights. This can be used to guide successive choice of sample sizes, and possible interventions to adjust successive kernel smoothing parameter values, and also to assess whether further refinement is likely to be additionally effective. Several illuminating examples appear in West (1992a).
1.3 Approximating mixtures by mixtures

Suppose the above adaptive strategy has \( n_0 = 500 \), so that \( g_1(\theta) \) is a mixture of 500 \( p \)-dimensional \( T \) distributions, and that the revision process stops here, \( g_1(\theta) \) being adopted as the ‘final’ importance sampling density to be used for Monte Carlo inference. It is straightforward to sample \( \Theta_1 = \{ \theta_{1,i}, \; i = 1, \ldots, n_1 \} \) from \( g_1(\theta) \) — the computational benefit of the components sharing a common scale matrix \( V_0 \) is apparent here. It is also straightforward to then evaluate the weights \( \Omega_1 = \{ w_{1,i}, \; i = 1, \ldots, n_1 \} \), though the denominator of \( w_{1,i} \) requires evaluation of the mixture \( g_1(\theta_{1,i}) \). The computational burden here clearly increases if further refinement of importance functions with rather larger sample sizes, is desired. One way of reducing such computations is to note that, quite typically, approximating \( p(\theta) \) using a mixtures of several thousand \( T \), or other, distributions is really overkill; even very irregular densities can be adequately matched using mixtures having far fewer components. The Monte Carlo kernel density construction, in particular, typically leads to a huge redundancy, with many points \( \theta_{0,i} \) closely grouped and contributing essentially similar components to the overall mixture density. In the context of raw density estimation, West (1990) discusses the reduction of kernel estimates to mixtures of much smaller numbers of components, often lower than 10% of the original sample size number, using a particular form of clustering. The discussion in West and Harrison (1989, Section 12.3), on issues and techniques involved in approximating mixtures generally, is also relevant.

A very basic method of ‘clustering’ mixture components, combining ideas from each of these two references, is used in West (1992a). At the simplest, it involves reducing the number of components by replacing ‘nearest neighbouring’ components with some form of average. The examples below, and those in West (1992a), involve reducing mixtures of \( n \) in several hundreds or thousands, to around 10% (though sometimes rather less) of the initial value, and perform this reduction using the following simple clustering routine (see West, 1992a, for further details).

1. Set \( r = n \), and, starting with the \( r = n \) component mixture (1), choose \( k < n \) as the number of components for the final, reduced mixture.
2. Sort the \( r \) values of \( \theta_{1,i} \) in \( \Theta \) in order of increasing values of weights \( w_j \) in \( \Omega \); thus \( \theta_1 \) corresponds to the component with smallest weight.
3. Find the index \( i, \; (i = 1, \ldots, r) \), such that \( \theta_i \) is the nearest neighbour of \( \theta_1 \), and reduce the sets \( \Theta \) and \( \Omega \) to sets of size \( r - 1 \) by removing components 1 and \( i \), and inserting ‘average’ values \( \theta_* \) = \( (w_1 \theta_1 + w_i \theta_i)/(w_1 + w_i) \) and \( w_1 = w_1 + w_i \). Set \( r = r - 1 \).
4. Proceed to (2), and repeat the procedure, stopping here only when \( r = k \).
5. The resulting mixture has the form (1) with \( r \) reduced to \( k \), the locations based on the final \( k \) averaged values, with associated combined weights, the same scale matrix \( V \) but new, and larger, window-width \( h \), based on the current, reduced ‘sample size’ \( r \) rather than \( n \).

The reduction process can be monitored, by evaluating and plotting margins of the mixture over fairly crude grids, at stages in the reduction process. If in reducing from, say, \( n \) to 80\% \( n \) components, these densities do not appear change much, then further reduction may proceed, and so forth in later stages of reduction. This is one possible route to approximating mixtures of large numbers of components with mixtures of far fewer. There are other possibilities, using alternative clustering algorithms, such as those and with foundation in Bayesian density estimation models (West, 1990), as illustrated by Müller (1993).

2. Sequential updating and dynamic models

2.1 Introduction

A central concept in Bayesian analysis is that of updating a prior to posterior distribution for a random quantity or parameter vector \( \theta \) based on received data summarised through a likelihood function for the parameter. With discrete approximations, this involves mapping a prior set of points and weights to a posterior set, possibly with some form of smoothing involved at both prior and posterior stages. These issues, and others, are sharply evident in sequential modelling of time series using dynamic models (West and Harrison, 1989), where the progressive revision of posterior and predictive distributions requires calculations that are typically impossible to perform exactly. Modellers have developed a variety of approaches to analytic and numerical approximation of such distributions, including direct quadrature methods (Kitagawa, 1987) and more efficient adaptive methods (Pole and West, 1990). The current paper extends the adaptive importance sampling techniques just reviewed to the dynamic modelling context to overcome the problems associated with quadrature techniques: notably, the need for ‘grids’ of evaluation points in parameter spaces.
to be changed as time progresses and as data indicate support for different regions in parameter space, the severe computational demands in problems with more than very few parameters, and the difficulties in reconstructing smooth posterior distributions based on approximate evaluation at only very few points in what may be several dimensions. These issues currently limit the development of quadrature techniques quite generally, outside the dynamic modelling framework, and have led investigators in other fields to turn to simulation based approaches of one form or another. Building on these approaches, the development in West (1992a) of adaptive importance sampling functions based on mixtures of standard distributions, and smooth reconstruction of posteriors using generalised kernel density estimates, may be naturally extended to the sequential dynamic modelling context.

2.2 Dynamic models

Interest lies in analysis of a dynamic model for a time series of (possibly multivariate) observations \( Y_t \), \( (t = 1, 2, \ldots) \), described as follows (West and Harrison, 1989, Section 13.5). At each time \( t \), \( Y_t \) has a known sampling distribution with density (or mass function) \( p_s(Y_t|\theta_t) \); here \( \theta_t \) is a \( p \)-vector of parameters, the state vector, and, conditional on \( \theta_t \), \( Y_t \) is assumed independent of \( Y_s \) and \( \theta_s \) for all past and future values of \( s \). The state vector is assumed to evolve in time according to a known, Markovian process described by an evolution density \( p_e(\theta_t|\theta_{t-1}) \); given \( \theta_{t-1}, \theta_t \) is conditionally independent of past values \( \theta_s \) for \( s < t - 1 \). Summarising the model, for each \( t = 1, 2, \ldots \), the defining equations are

\[
\begin{align*}
\text{Observation model :} & \quad (Y_t|\theta_t) \sim p_o(Y_t|\theta_t) \\
\text{Evolution model :} & \quad (\theta_t|\theta_{t-1}) \sim p_e(\theta_t|\theta_{t-1})
\end{align*}
\]

The sampling, or observation, density is suffixed ‘\( o \)’, and the evolution density is suffixed ‘\( e \)’, in order to identify them in the discussion below where ‘\( p \)’ is used generically to denote densities. Note that these densities may depend also on \( t \) and on other available historical information, though the notation ignores this for clarity. At each time \( t \), available information, including the observed values of \( Y_t, Y_{t-1}, \text{ etc.} \), is denoted \( D_t \). Currently, at time \( t - 1 \) prior to evolution through (3), historical information is summarised through (an approximation to) the posterior \( p(\theta_{t-1}|D_{t-1}) \). The primary computational problems addressed are:

(a) **Evolution step**: compute the current prior for \( \theta_t \), defined via

\[
p(\theta_t|D_{t-1}) = \int p_e(\theta_t|\theta_{t-1})p(\theta_{t-1}|D_{t-1})d\theta_{t-1};
\]

(b) **Updating step**: on observing \( Y_t \), compute the current posterior

\[
p(\theta_t|D_t) \propto p(\theta_t|D_{t-1})p_o(Y_t|\theta_t).
\]

Subsidiary calculations include forecasting ahead (discussed below) and filtering problems for retrospection (not considered in this paper). It is important to note that analysis proceeds sequentially over time, and it is explicitly recognised that the summary posterior of historical information, namely \( p(\theta_{t-1}|D_{t-1}) \), is the only information relevant to the past that is currently (at time \( t - 1 \)) available for further analysis, and this must be borne in mind when developing numerical (or any other) approximations to dynamic model analyses. In addition, the process is to be repeated as time progresses, so the form of any approximation to the posterior after the updating step must be such that, moving from time \( t \) to \( t + 1 \) when this posterior becomes the input to the evolution/updating analysis for \( \theta_{t+1} \), the required calculations can proceed similarly. In developing Monte Carlo based approximations (or any other numerical approach) the objective is therefore to provide structured information about \( (\theta_t|D_t) \), after the updating step, that has the same qualitative form as that input for \( (\theta_{t-1}|D_{t-1}) \). With reference to development above, a Monte Carlo/importance sampling approximation to any distribution reduces to summary components given by the four elements \( \Gamma \), so that, with such an approach to analysis, the entire evolution/updating process at each time can be viewed as one of appropriately updating such components, with subsidiary computations for forecasting, etc. This directly parallels the usual, analytic process in linear, normal models, (West and Harrison, 1989), and the numerical process using quadrature of Pole and West (1990).
Suppose that \( p(\theta_{t-1}|D_{t-1}) \) has been previously approximated via \( \Gamma_{t-1} = \{ g_{t-1}, n_{t-1}, \Theta_{t-1}, \Omega_{t-1} \} \), where \( g_{t-1}(\theta_{t-1}) \) is a (final) importance density used for inference about \( (\theta_{t-1}|D_{t-1}) \), \( n_{t-1} \) is the (final) Monte Carlo sample size in that inference, \( \Theta_{t-1} = \{ \theta_{t-1,i} \}, \ i = 1, \ldots, n_{t-1} \) is the sample from \( g_{t-1}(\theta_{t-1}) \), and \( \Omega_{t-1} = \{ w_{t-1,i} \}, \ i = 1, \ldots, n_{t-1} \) the associated weights. The objective is to perform the evolution/ updating and forecasting computations and finally summarise \( p(\theta_{t}|D_{t}) \) in terms of \( \Gamma_{t} = \{ g_{t}, n_{t}, \Theta_{t}, \Omega_{t} \} \), and this may be done as follows.

### 2.3 Computation: evolution step

The following facts are of use in computations for the evolution step.

(a) Various features of the prior \( p(\theta_{t}|D_{t-1}) \) of interest can be computed directly using the Monte Carlo structure \( \Gamma_{t-1} \). The prior mean, for example, is computable as

\[
E[\theta_{t}|D_{t-1}] \approx \sum_{i=1}^{n_{t-1}} w_{t-1,i} E[\theta_{t}|\theta_{t-1,i}],
\]

where \( E_{\theta} \) stands for expectation under the evolution distribution \( p_{\theta} \), if this expectation is available in closed form.

(b) Similarly, the prior density function can be evaluated by Monte Carlo integration at any required point, viz

\[
p(\theta_{t}|D_{t-1}) \approx \sum_{i=1}^{n_{t-1}} w_{t-1,i} p_{\theta}(\theta_{t}|\theta_{t-1,i}). \tag{6}
\]

(c) An initial Monte Carlo sample of size \( n_{t-1} \) may be drawn from the prior by generating one value of \( \theta_{t} \) from \( p_{\theta}(\theta_{t}|\theta_{t-1,i}) \), for each \( i = 1, \ldots, n_{t-1} \). The resulting sample points, denoted \( \Theta_{t} \), provide, at the very least, starting values for the evaluation of the prior as in (b) above, for display purposes, and, importantly, regions of \( \theta \) parameter space to concentrate on in beginning the updating step, below.

(d) This prior sample \( \Theta_{t} \) may be used with weights \( \Omega_{t-1} \) to construct a generalised kernel density estimate of the prior. In many models, this is of little interest since the prior may be evaluated as in (b) above (unless the evolution density is extremely complex and difficult to work with.) Consider, however, a model in which the evolution equation is degenerate, \( \theta_{t} = G_{t}(\theta_{t-1}) \) with probability one, for some known, possibly non-linear and time-dependent function \( G_{t} \). This covers many interesting examples, such as that appearing in Section 4.4 below. Then prior evaluation as in (b) is vacuous, and so, since values of the prior are required as input to Bayes’ theorem (5) in the updating step, some form of interpolation/smoothing is needed.

(e) Note, at this point, that subsidiary computations for forecasting ahead can be performed along these lines. Forecasting \( (Y_{t}|D_{t-1}) \), for example, requires computing \( \int p_{\theta}(Y_{t}|\theta_{t}) p(\theta_{t}|D_{t-1}) d\theta_{t} \). The observation density typically has a standard form, so that Monte Carlo computations can be performed to approximate forecast moments and probabilities. For example, the forecast mean is evaluated as

\[
E[Y_{t}|D_{t-1}] \approx \sum_{\theta_{t} \in \Theta_{t}} w_{t-1,i} E_{\theta_{t}}[Y_{t}|\theta_{t}],
\]

where \( E_{\theta_{t}} \) stands for expectation under the observation distribution \( p_{\theta} \), assuming this expectation is available in closed form. Forecast probabilities are similarly derived. By simulating from the observation density for each value of \( \theta_{t} \in \Theta_{t} \), and using the associated weights \( \Omega_{t-1} \), regions of interest in \( Y_{t} \) space can be identified, and the predictive density/mass function evaluated there. Similar comments apply to forecasting more than one step ahead; the process of simulating from the evolution density is repeated into the future, generating sample of future parameter vectors, and proceeding to inference about the future values of the series. This too is illustrated below.

It is important to note the generality of the above strategy for computations. At no stage is it necessary, or interesting, to worry about functional forms of evolution equations, or to cater for many special cases, as alternative approaches, such as using quadrature in Pole and West (1990), must. This simplifies programming the analysis; all that is needed is a collection of general routines for evaluating and sampling from the evolution and observation densities, and, if required, generating kernel estimates.
2.4 Computations: updating step

Following evolution, the prior \( p(\theta_1|D_{t-1}) \) to input Bayes’ theorem (5) for updating is available in approximate form either via evaluation in (6) at any desired points, or in terms of a generalised kernel form as described in (d) of the previous section. Updating may then proceed using adaptive Monte Carlo density estimation of Section 2. This begins with an initial ‘guess’ at the posterior \( p(\theta_1|D_{t-1}) \), denoted \( g_{t,0}(\theta_1) \), to be used as an initial importance sampling function. Choice of this density depends on context, though may be guided by general ideas described in the examples below. This provides the starting point for application of the adaptive strategy of section 1. On completion, this results in a final summary of the posterior given by the quadruple \( \Gamma_1 = \{g_{t}, n_t, \Theta_t, \Omega_t\} \), where \( g_t(\theta_1) \) is the final importance sampling density for the posterior \( p(\theta_1|D_{t}) \), \( n_t \) is the final Monte Carlo sample size, and \( \Theta_t \) is the set of \( n_t \) points in parameter space at which the Monte Carlo weights in \( \Omega_t \) are evaluated.

3. EXAMPLES

Example 1

The first example concerns a familiar and simple model, chosen so that exact calculations can be performed to provide comparison with approximation based on the above strategy. The model is a normal, linear, first-order polynomial model (West and Harrison, 1989, Chapter 2), in which \( p = 1 \), so that \( \theta_t = \theta_1 \) is scalar, the observation density is normal with mean \( \theta_t \) and unit variance, \( (Y_t|\theta_t) \sim N(\theta_t, 1) \), and the evolution density is normal, \( (\theta_t|\theta_{t-1}) \sim N(\theta_{t-1}, 1) \). If it is assumed, initially, that \( (\theta_1|D_0) \) is normal, then the standard analysis applies to give easy calculation of prior, posterior and forecast distributions for any time \( t \), such distributions being, of course, normal. The example here uses a series of length 100 generated from this model with \( \theta_0 = 0 \). The numerical analysis is structured as follow: (i) The initial prior is standard \( T \) with 9 degrees of freedom; (ii) in each updating step, computing the prior densities \( p(\theta_1|D_{t-1}) \) for evaluation of the (unnormalised) posteriors is done via equation (6); (iii) updating uses adaptive Monte Carlo density estimation, the sample size of all final approximations set at 1500 – the first approximation \( g_{t,0}(\theta_1) \) at each stage generated as follows. The summary \( \Gamma_{t-1} \) for \( (\theta_{t-1}|D_{t-1}) \) (with sample size \( n_{t-1} = 1500 \)) is used to generate an initial sample \( \Theta_{t}^* = \{\theta_{t,1}^*, \ldots, \theta_{t,n_t}^*\} \) for \( \theta_t \), the estimate of mean and variance given by \( \mu_t = \frac{\sum_{i=1}^{n_t} w_{t-1,i} \theta_{t,i}^*}{\sum_{i=1}^{n_t} w_{t-1,i}} \), and \( \Sigma_t = \frac{\sum_{i=1}^{n_t} w_{t-1,i} (\theta_{t,i}^* - \mu_t)^2}{\sum_{i=1}^{n_t} w_{t-1,i}} \), respectively. Clustering reduces this \( n_{t,0} = 100 \) components \( \Theta_{t,0} = \{\theta_{t,0,j}, j = 1, \ldots, 100\} \), with reduced weights \( \Omega_{t,0} = \{w_{t,0,j}, j = 1, \ldots, 100\} \). Then \( g_{t,0}(\theta_1) \) is taken as the mixture (1) using these reduced weights and points, the variance \( V = 16\Sigma_t \), window-width \( h \) chosen as in Silverman (1986) but using a \( T_0 \) kernel. The factor 16 inflates the spread to provide a conservative initial guess. The adaptive strategy now proceeds with \( n_{t,0} = 250, n_{t+1} = 350 \), and, finally, \( n_t = n_{t+2} = 1500 \). After sampling and computing moments in the first two stages, the components, of sizes 250 and 350 respectively, are reduced to 100 by clustering before sampling for the next iteration.

Inferences at each stage are based on the final, full Monte Carlo summaries of size 1500. Some posterior densities at times \( t = 10, 20, \ldots, 90 \) appear in Figure 1, with the current, actual value of \( \theta_t \) appearing on the axis as ‘X’, the corresponding datum \( Y_t \) as ‘Y’. In fact each frame plots two densities, the second being the exact, normal posterior based on the usual normal theory analysis. In none of these plots, and indeed over all time \( t = 1, \ldots, 100 \), is there a meaningful difference between the exact and numerical approximation.

Example 2

A similar example, though not normal, is analysed in West (1992b) – the normality of evolution and observation errors replaced with Student T distributions. Such models have been popular in the literature (Kitagawa, 1987; Pole and West, 1990, and references therein; West, 1981). This permits larger ‘jumps’ in the level \( \theta_t \) of the time series \( Y_t \), and also accommodates outlying observations. West (1992b) demonstrates the efficacy of the adaptive mixture techniques in analysing such a model. Unlike the normal case above, the posteriors for level parameters can become heavily skewed and even bimodal, evidencing occurrences of very extreme observations that induce ‘conflict’ between priors and likelihood functions. The facilities for outlier rejection and adaptation to level changes in heavy-tailed models are illustrated in West (1992b) using the numerical methods described here.
Figure 1. Posterior densities for normal process

Example 3

A simplified version of a model of Andrade Netto et al. (1978), discussed in Kitagawa (1978), illustrates uni-parameter analyses further, with some interesting features. The model is

\[ (Y_t|\theta_1) \sim N[0.05\theta_1^2, 10], \]
\[ (\theta_1|\theta_{t-1}) \sim N[0.5\theta_{t-1} + 25\theta_{t-1}/(1 + \theta_{t-1}^2), 1]. \]

The evolution structure here is bifurcating; the conditional mean $E_{\theta_1}[\theta_1|\theta_{t-1}]$ near $\theta_{t-1} = 0$ is approximately linear with gradient 25.5, so that small values of $\theta_{t-1}$ tend to evolve to very much larger values of $\theta_t$; as a result, very regular posteriors $p(\theta_{t-1}|D_{t-1})$, unimodal near zero, evolve into bimodal priors $p(\theta_t|D_{t-1})$, the mass being pushed away from the origin through the deterministic component of the evolution equation. The observation equation leads to further irregularities. For negative observed values of $Y_t$, the likelihood function, proportional to $p(Y_t|\theta_t)$ is unimodal at zero, and symmetric about its mode; for positive observations, however, the likelihood has an antimode at zero, and is symmetric about zero with modes at $\pm \sqrt{30Y_t}$.

The analysis for just three observations generated from this model is summarised here. The process is simulated from $\theta_0 = 0$, the three simulated $\theta_t$ values, and the corresponding simulated $Y_t$ values, noted on the lower axes of the frames of Figure 2. Analysis is as follows. (i) The initial prior for $(\theta_1|D_0)$ is Student $T$ on 9 degrees of freedom, with mode at 0 and scale factor 100; thus $\theta_1/10$ is standard $T_9$. (ii) Values of the prior densities $p(\theta_1|D_{t-1})$, required for evaluating the (unnormalised) posteriors in Bayes’ theorem (5), are calculated using the exact evolution density in the Monte Carlo expression (6). (iii) Updating uses adaptive Monte Carlo density estimation, the sample size of all final approximations set at 2000. The first
approximation $g_{t,0}(\theta_t)$ at each stage being generated as follows. First, $\Gamma_{t-1}$ (with sample size $n_{t-1} = 2000$) is used to generate an initial sample $\Theta^*_t = \{\theta^*_i \mid i = 1, \ldots, n_{t-1}\}$ for $\theta_t$. Clustering then applies to reduce this from 2000 to $n_{t,0} = 100$ components $\Theta_{t,0} = \{\theta_{t,0,j} \mid j = 1, \ldots, 100\}$, with associated reduced weights $\Omega_{t,0} = \{w_{t,0,j} \mid j = 1, \ldots, 100\}$. Then $g_{t,0}(\theta_t)$ is taken as the mixture (1) using these reduced weights and points, the variance $V = R_t$, and the kernel is again $T_0$. The adaptive strategy now proceeds using $n_{1,0} = 250$, $n_1 = 500$ and, finally, $n_2 = n_{2,2} = 2000$. Finally, the components, of sizes 250 and 500 respectively, are reduced to 100 before sampling for the next iteration.

**Figure 2. Inference in bifurcating model**

For $t = 1, 2$ and 3, the evaluated prior $p(\theta_t | D_{t-1})$, and the likelihood function proportional to $p_y(Y_t | \theta_t)$ (the latter standardised to unity at the maximum), appear in the first two frames of each row of graphs in Figure 2. The third frame in each row displays, as a broken line, the final density $g_t(\theta_t)$, a mixture of 100 $T_0$ densities. As a full line in the final frames, the ‘exact’ posterior based on the displayed prior and likelihood function is graphed. This is computed as the approximately normalised product of the prior and likelihood, with evaluations made over a fine grid of 300 points across the displayed interval of ±20. In the cases
of extreme bimodality, the importance density is naturally less peaked at the modes, though probabilities computed under the two densities are substantially similar. Continuing analysis to \( t = 4, 5, \ldots \), this pattern of behaviour is reproduced, with the two densities in the third frame essentially coinciding apart from cases of extreme bimodality when minor differences appear around the modes. Quantified inferences are based on the actual Monte Carlo approximations \( \Gamma_t \), rather than by using the importance densities as approximations directly, so that the under-estimation of density heights near peaks is irrelevant. As importance sampling densities for the displayed 'exact' posteriors, the displayed forms are excellent. Very similar results are achieved by repeating the analysis with larger sample sizes in the Monte Carlo. With the adaptive kernel density estimation used here, samples of 2000 at the final stage are typically quite adequate.

As final comment, it may be shown that, in fact, the prior and posterior distributions in the example are theoretically guaranteed to be symmetric about zero, for all time \( t \), when the initial prior \( p(\theta_t | D_0) \) is itself symmetric about zero (as is the case in the above analysis.) None of the estimated densities in Figure 2 departs significantly from symmetry about zero, although, due to the randomness inherent in Monte Carlo approaches, small deviations away from symmetry are apparent. It should be noted that knowledge of such a feature, or any other theoretically based information about form, may be incorporated into the analysis if required. The symmetry about zero here could easily be incorporated by taking absolute values of all sampled \( \theta_t \), and then reflecting resulting samples, and smoothed densities, about the origin.

**Example 4**

A final example revisits a class of models used in assessing the effect of television advertising on the 'awareness' of consumer populations, introduced in Migon and Harrison (1985), and discussed in West and Harrison (1989, Section 14.4). The particular model chosen for illustration here is one in which prior evaluations using (6) are impossible since the evolution distribution is degenerate. This example thus provides (a) a multi-parameter model, with \( p = 3 \); (b) an opportunity to explore numerical analysis in this familiar model class, and establish a basis for possible future comparison with previous analytic approximations; and, (c) a case in which direct generalised kernel estimation is used to sequentially reconstruct posterior distributions from Monte Carlo analyses.

Much detail on the models and context of the advertising problem appears in the above references, to which the reader may refer, although the data analysis here is somewhat different to previous analyses. Raw 'advertising awareness' data, in the form of binomial counts, are collected over a period of 56 weeks, together with an associated regressor variable 'TVR' measuring weekly advertising expenditure. Indexing weeks by \( k \), \( k = 1, \ldots, 56 \), let \( Z_k \) denote recorded counts in week \( k \), and \( X_t > 0 \) the 'TVR'. The model assumes \( Z_k \) to be conditionally independent with binomial probabilities \( \pi_k \) defined (using the notation of West and Harrison, 1989, Section 14.4) via

\[
\pi_k = 0.1 + E_{k},
\]

\[
E_{k} = 0.7 - (0.7 - \rho E_{k-1}) \exp(-\kappa X_t),
\]

for \( k = 1, \ldots, 56 \). Quantities \( \rho \) and \( \kappa \), and the initial effect \( E_0 \), are initially uncertain. The data are the first 56 weeks of series in Table 14.1, Section 14.4 of West and Harrison (1989). Analysis here groups the raw weekly data into seven, 8-week periods, defining data \( Y_t \) via

\[
Y_t = \{ Z_{8t-7}, Z_{8t-6}, \ldots, Z_{8t} \}, \quad (t = 1, \ldots, 7).
\]

For each \( t \), the three uncertain parameters are \( \rho, \kappa \) and \( E_{8t-8} \), the effect \( E_{8t-8} \) being that in the final week of the previous 8-week period. In the first 8-week period, for example, \( t = 1 \) so that \( E_{8t-8} = E_0 \), the current effect just prior to starting the advertising campaign. The sampling density in period \( t \) is the product of 8 component binomials, \( \prod_{k=8t-7}^{8t} p(Z_k | \pi_k) \), in which the \( \pi_k \) can be evaluated as functions of \( \rho, \kappa \), and \( E_{8t-8} \) by repeat application of equations (7). In evolution to times \( t = 2, \ldots, 7 \), \( \rho \) and \( \kappa \) remain unchanged, and the third parameter \( E_{8t-8} \) is a deterministic, non-linear function of that previously, \( E_{8t-16} \); given \( E_{8t-16} \), evolution calculations again just involve repeat application of the second equation in (7) to evaluate \( E_{8t-8} \). Note that the parameters are restricted by inequalities \( 1 > \rho > 0 \), \( \kappa > 0 \) and \( 0.7 > E_{8t-8} > 0 \).

Since importance sampling densities are to be constructed as mixtures of tri-variate T distributions, the restricted parameters \( \rho, \kappa \) and \( E_{8t-8} \) are transformed to real-valued versions \( \theta_t = (\theta_1, \theta_2, \theta_3)^T \) where

\[
\theta_1 = \log \rho, \quad \theta_2 = \log \kappa, \quad \theta_3 = \log E_{8t-8}.
\]
\( \theta_1 = \log(\rho/(1-\rho)), \theta_2 = \log(\kappa) \) and \( \theta_{1..3} = \log((E_{8..8}E_{8..8})/(0.7-E_{8..8})) \). The analysis summarised here (validated by repeat analysis using different, and larger, Monte-Carlo sample sizes at each stage) has the following features, again essentially similar in structure to previous examples.

(i) The initial prior for \( \theta_1 = (\rho, \kappa, E_0)^\prime \) given \( D_0 \) is trivariate T with 9 degrees of freedom; the prior mean is \( (2.0,-3.5,-0.75)^\prime \), and the prior scale factors are \( (0.20, 0.25, 0.25)^\prime \), so \( V[\theta_1|D_0] = 0.20(9/7) \) and \( V[\theta_2|D_0] = V[\theta_3|D_0] = 0.25(9/7) \). The data set used here is taken from Section 14.4 of West and Harrison (1989), and this prior specification is consistent with that of the analysis there. The prior has no non-zero correlations.

(ii) The Monte Carlo summary \( \Gamma_{t-1} \) of \( p(\theta_{t-1}|D_{t-1}) \) evolves by just mapping the third elements of vectors in \( \Theta_{t-1} = \{ \theta_{t-1,i}, i = 1, \ldots, n_{t-1} \} \). Thus \( \Gamma_{t-1} \) evolves to the prior summary \( \Gamma^* \) via just this, deterministic transformation. Values of the prior \( p(\theta_1|D_{t-1}) \) required for evaluating the (unnormalised) posteriors in Bayes’ theorem (5), are calculated from a kernel density reconstructed version, based on this evolved Monte Carlo approximation \( \Gamma^*_t \) and using trivariate \( T_0 \) kernels.

(iii) Updating uses adaptive Monte Carlo density estimation, the sample size of all final approximations set at 4000. The first approximation \( g_{t,0}(\theta_1) \) at each stage is the weighted kernel estimate based on Monte Carlo summary \( \Gamma^*_t \) but with the variance matrix scaling the kernels inflated by a factor of 4; this has the same location characteristics as the reconstructed prior \( p(\theta_1|D_{t-1}) \), but greater spread. The process of adaptive refinement proceeds through two stages, as in the examples above, with sample sizes successively increased from 1000, 1500 and then the final 4000. Between stages, the clustering technique of Section 2.4 is applied to reduce the number of components (from 1000 and 1500 respectively) to 500.

At the end of analysis at each stage \( t = 1, \ldots, 7 \), the final summary \( \Gamma_t \) of \( p(\theta_1|D_t) \) is evolved to \( \Gamma^*_{t+1} \) and further evolution over the \( (t+1)^{st} \) 8-week period is performed to produce step-ahead forecasts of the probabilities \( \pi_k \) in (7); thus, at time \( t \), direct transformation of \( \Gamma_t \) leads to Monte Carlo predictive distributions for \( \pi_k \) for each \( k = 8t - 7, 8t - 6, \ldots, 8t \). Figure 3(b) displays the TVR inputs \( X_k \) versus time \( k \). In Figure 3(a), the raw data proportions \( Z_k/66 \) are indicated at each time \( k \), joined up over time with a
full line. The step-ahead forecast distributions for $\pi_k$, generated at the end of each 8-week period (marked by vertical dotted lines) for the coming 8 weeks, are indicated in terms of medians (the small circles) and vertical lines representing 90% equal-tails intervals. Note that data variation about the $\pi_k$ is binomial; these intervals are for the $\pi_k$ directly, so forecast intervals for the $Z_k$ would be similarly located but wider due to the additional binomial variation. At these levels of $\pi_k$, around 0.3-0.4, binomial standard deviations are around 0.06. The analysis compares well with similar displays (albeit from a rather different model and analysis) in West and Harrison (1989, Section 14.4).

With the time independence of the parameters $\rho$ and $\kappa$, the data could alternatively be analysed as one, rather than sequentially. Here, to illustrate the techniques in sequential modelling, it has been assumed that summary inferences are required after each 8-week period, and that this summary at each stage is all that is carried forward to the next; previous data, and therefore any opportunity to perform revised analyses using all data so far, is assumed lost at the end of each stage.

However, to provide comparison, such a one-shot analysis was performed, using all the data at once. This analysis, along the lines of that in West (1992a), is simpler to perform, and is not subject to the possibility of approximation errors ‘building-up’ through the sequential evolution/updating process. It is based on the same prior specification as above, and uses adaptive refinement of importance sampling functions through stages of sample sizes of 1000, 1500, 1500 and, finally, 4000. Again, at each stage, reduction is made to 500 components before proceeding. The resulting agreement with the sequential analysis is excellent. Final posterior means and standard deviations differ only in the third decimal place in each case, as does the estimated correlation between the transforms of $\rho$ and $\kappa$. Correlation between the transforms of $\rho$ and $E_{0\theta}$ is estimated, to two decimal places, at 0.94 rather than 0.95, that between the transforms of $\kappa$ and $E_{0\theta}$ at $-0.73$ rather than $-0.74$. These are differences of negligible practical import and easily attributable to the inherent randomness of the Monte Carlo.

References


