Combined parameter and state estimation in simulation-based filtering

Jane Liu & Mike West

Abstract

We discuss simulation-based sequential analysis – or particle filtering – in dynamic models, with a focus on sequential Bayesian learning about time-varying state vectors and fixed model parameters simultaneously. This includes a general approach that combines old ideas of smoothing using kernel methods with newer ideas of auxiliary particle filtering. We highlight specific smoothing approaches that have interpretation as adding "artificial evolution noise" to fixed model parameters at each time point to address problems of sample attrition and prior:data conflict. We introduce a new approach that permits such smoothing and regeneration of sample points of model parameters without the "loss of historical information" inherent in earlier methods; this is achieved using shrinkage modifications of kernel smoothing, as introduced by the second author in the early 1990s. Following some theoretical development, discussion and a small simulated example to demonstrate its efficacy, we report some experiences in using the method in a challenging application in multivariate dynamic factor models for financial time series, as recently studied using MCMC methods by authors, their collaborators, and other researchers. Some summary comments and comparisons with MCMC methods are given in this applied context, and we conclude with some discussion of general issues of practical relevance, and suggestions for further algorithmic development.

1 Introduction and Historical Perspective

Much of the recent and current interest in simulation-based methods of sequential Bayesian analysis of dynamic models has been focused on improved methods of filtering for time-varying state vectors. We now have quite effective algorithms for time-varying states, as represented throughout this volume. Variants of the auxiliary particle filtering algorithm (Pitt and Shephard 1999b), in particular, are of proven applied efficacy in quite elaborate models. However, the need for more general algorithms that deal simultaneously with both fixed model parameters and state variables is practically pressing. We simply do not have access to efficient and effective methods of treating this

problem, especially in models with realistically large numbers of fixed model parameters. It is a very challenging problem.

A short historical commentary will be of interest in leading into the developments presented in this chapter. In the statistics literature per se, simulation-based filtering can be seen as a natural outgrowth of converging interests in the late 1980s and early 1990s. For several decades researchers involved in sequential analysis of dynamic models, in both statistics and various engineering fields, have been using discrete numerical approximations to sequentially updated posterior distributions in various "mixture modelling" frameworks. This literature has involved methods for both time-evolving states and fixed parameters, and is exemplified by the important class of adaptive multi-process models used in Bayesian forecasting since the early 1970s (Harrison and Stevens 1976, Smith and West 1983, West and Harrison 1997). During the 1980s, this naturally led to larger-scale analyses using discrete grids of parameter values, though the combinatorial "explosion" of grid sizes with increasing parameter dimension limited this line of development. Novel methods using efficient quadrature-based, adaptive numerical integration ideas were introduced in the later 1980s (Pole 1988, Pole, West and Harrison 1988, Pole and West 1990). This involved useful methods in which discrete grids – of both fixed model parameters and state variables – themselves change over time as data is processed, sequentially adapting the discrete posterior approximations by generating new "samples" as well as associated "weights." This work recognised the utility of the Markov evolution equations of dynamic models in connection with the generation of new grids of values for time-evolving state variables. It similarly recognised and addressed the practically critical issues of "diminishing weights" on unchanging grids of parameter values, and the associated need for some method of interpolation and smoothing to "regenerate" grids of values for fixed model parameters. In these respects, this adaptive deterministic approach naturally anticipated future developments of simulation-based approaches. Again, however, parameter dimension limited the broader applicability of such approaches.

The end of the 1980s saw a developing interest in simulation-based methods. Parallel developments in the early 1990s eventually led to publication of different but related approaches (West 1993a, West 1993b, Gordon, Salmond and Smith 1993). All such approaches involve methods of evolving and updating discrete sets of sampled state vectors, and the associated weights on such sampled values, as they evolve in time. It has become standard to refer to sampled values as "particles." The above referenced works, and others, have again highlighted the utility of the convolution structure of Markov evolution equations for state variables in generating Monte Carlo samples of time-evolving states. Most approaches recognise and attempt to address the inherent problems of degrading approximation accuracy as particulate representations are updated over time – the issues of particle "attrition" in resampling methods and of "weight degeneracy" in reweighting methods. These

issues are particularly acute in approaches that aim to deal with fixed model parameters as part of an extended state vector. This was addressed in the approaches of West (1993a,b). Openly recognising the need for some kind of interpolation/smoothing of "old" parameter particles to generate new values, this author used local smoothing based on modified kernel density estimation methods, the modifications being based on Bayesian reasoning and geared towards adjusting for the over-smoothing problems inherent in standard kernel methods. In later years, this approach has been used and elaborated. For example, it has been extended to include variable shapes of multivariate kernels to reflect changing patterns of dependencies among parameters in different regions of parameter space (Givens and Raftery 1996), as explicitly anticipated in West (1993a,b). A related approach may be referred to as the "artificial evolution" method for model parameters. This relates to the work of Gordon et al (1993), who introduced the idea of adding additional random disturbances – or "roughening penalities" – to sampled state vectors in an attempt to deal with the degeneracy issues. Extending this idea to fixed model parameters leads to a synthetic method of generating new sample points for parameters via the convolution implied by this "artificial evolution." This neat, ad-hoc idea is easily implementable, but suffers the obvious drawback that it "throws away" information about parameters in assuming them to be time-varying when they are, in fact, fixed. The same drawback arises in using the idea in its original form for dynamic states.

In this current chapter, we take as our starting point these methods of dealing with fixed model parameters, and address the issues of reconciling them and then embedding a generalised algorithm within a sequential auxiliary particle filtering context. Section 2 discusses particle filtering for state variables, and describes the general framework for combined filtering on parameters and state variables. Section 3 focuses on the problems arising in simulation-based filtering for parameters, reviews the kernel and artificial evolution methods, identifies the inherent structural similarities in these methods, and then introduces a modified and easily implemented approach that improves upon both by resolving the problem of information loss they each imply. Returning to the general framework in Section 4, we describe a general algorithm that extends the more-or-less standard auxiliary particle filtering approach for state variables to include model parameters. We give a simple example for illustration, and then, in Section 5, report on some experiences in using the method in a challenging application in multivariate dynamic factor models for financial time series. Section 6 concludes with some summary comments, discussion and suggestions for new research directions.

2 General Framework

2.1 Dynamic Model and Analysis Perspective

Assume a Markovian dynamic model for sequentially observed data vectors $\boldsymbol{y}_t$, $(t = 1, 2, \ldots)$, in which the state vector at time t is $\boldsymbol{x}_t$ and the fixed parameter vector is $\boldsymbol{\theta}$. The model is specified at each time t by the observation equation defining the observation density

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}) \tag{2.1}$$

and the Markovian evolution equation, or state equation, defining the transition density

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}). \tag{2.2}$$

Each $\boldsymbol{y}_t$ is conditionally independent of past states and observations given the current state $\boldsymbol{x}_t$ and the parameter $\boldsymbol{\theta}$, and $\boldsymbol{x}_t$ is conditionally independent of past states and observations given $\boldsymbol{x}_{t-1}$ and $\boldsymbol{\theta}$. This covers a very broad class of practically interesting models (West and Harrison 1997).

Sequential Monte Carlo methods aim to sequentially update Monte Carlo sample approximations to the sequences of posterior distributions $p(\boldsymbol{x}_t, \boldsymbol{\theta}|D_t)$ where $D_t = \{D_{t-1}, \boldsymbol{y}_t\}$ is the information set at time t. Thus, at time t this posterior is represented by a discrete sample of points and weights (the latter possibly though not necessarily uniform weights). On observing the new observation $\boldsymbol{y}_{t+1}$ it is desired to produce a sample from the "current" posterior $p(\boldsymbol{x}_{t+1}, \boldsymbol{\theta}|D_{t+1})$. There have been numerous contributors to theoretical and applied aspects of research in this area in recent years (West 1993b, Gordon et al. 1993, Kitagawa 1998, Liu and Chen 1995, Berzuini, Best, Gilks and Larizza 1997, Pitt and Shephard 1999b) and the field has really grown dramatically both within statistical sciences and in related fields including, especially, various branches of engineering (Doucet 1998). The current volume represents a comprehensive catalogue of recent and current work.

The most effective methods all utilise the state equation to generate sample values of the current state vector $\boldsymbol{x}_{t+1}$ based on past sampled values of the state $\boldsymbol{x}_t$. This is critical to the utility and performance of discrete approximation methods, as the generation of new sets of states from what is usually a continuous state transition density allows the posterior approximations to "move around" in the state space as the state evolves and new data are processed. We focus now exclusively on auxiliary particle filters as developed in Pitt and Shephard (1999b), variants of which are, in our opinion, the most effective methods currently available – though that may change as the field evolves.

2.2 Filtering for States

Consider a model with no fixed parameters, or in which $\boldsymbol{\theta}$ is assumed known, so that the focus is entirely on filtering for the state vector. Standing at time t, suppose we have a sample of current states $\{\boldsymbol{x}_t^{(1)}, ..., \boldsymbol{x}_t^{(N)}\}$ and associated weights $\{\omega_t^{(1)}, ..., \omega_t^{(N)}\}$ that together represent a Monte Carlo importance sample approximation to the posterior $p(\boldsymbol{x}_t|D_t)$. This includes, of course, the special case of equal weights in which we have a direct posterior sample. Time evolves to $t+1$, we observe $\boldsymbol{y}_{t+1}$, and want to generate a sample from the posterior $p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_{t+1})$. Theoretically,

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_{t+1}) \propto p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_t) \tag{2.3}$$

where $p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_t)$ is the prior density of $\boldsymbol{x}_{t+1}$ and $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ is the likelihood function. The second term here – the prior for the state at time $t+1$ – is implied by the state equation as $\int p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{D}_t)d\boldsymbol{x}_t$. Under the Monte Carlo approximation to $p(\boldsymbol{x}_t|\boldsymbol{D}_t)$, this integral is replaced by a weighted summation over the sample points $\boldsymbol{x}_t^{(k)}$, so that the required update in equation (2.3) becomes

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_{t+1}) \propto p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}) \sum_{k=1}^{N} \omega_t^{(k)} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)}). \tag{2.4}$$

To generate Monte Carlo approximations to this density, an old and natural idea is to sample from $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)})$ for $k = 1, ..., N$, evaluate the corresponding values of the weighted likelihood function $\omega_t^{(k)} p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ at each draw, and then use the normalised weights as the new weights of the samples. This basic "particle filter" is an importance sampling method closely related to those of West (1993a,b). A key problem is that the sampled points come from the current prior of $\boldsymbol{x}_{t+1}$ and the resulting weights may be very small on many points in cases of meaningful separation of the prior and the likelihood function based on $\boldsymbol{y}_{t+1}$. West (1993a,b) developed an effective method of adaptive importance sampling to address this. The idea of auxiliary particle filtering (see Pitt and Shephard, 1999b, and the chapter by these authors in this volume) is similar in spirit but has real computational advantages; this works as follows. Incorporate the likelihood function $p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$ under the summation in equation (2.4) to give

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{D}_{t+1}) \propto \sum_{k=1}^{N} \omega_t^{(k)} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)}) p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1})$$

and generate samples of the current state as follows. For each $k = 1, \ldots, N$, select an estimate $\boldsymbol{\mu}_{t+1}^{(k)}$ of $\boldsymbol{x}_{t+1}$, such as the mean or mode of $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(k)})$. Evaluate the weights $g_{t+1}^{(k)} \propto \omega_t^{(k)} p(\boldsymbol{y}_{t+1}|\mu_{t+1}^{(k)})$. A large value of $g_{t+1}^{(k)}$ indicates that $\boldsymbol{x}_t^{(k)}$, when "evolving" to time $t+1$, is likely to be more consistent

with the datum $\boldsymbol{y}_{t+1}$ than otherwise. Then indicators j are sampled with probabilities proportional to $g_{t+1}^{(j)}$, and values $\boldsymbol{x}_{t+1}^{(j)}$ of the current state are drawn from $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(j)})$ based on these "auxiliary" indicators. These sampled states are essentially importance samples from the time $t+1$ posterior and have associated weights

$$\omega_{t+1}^{(j)} = \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}^{(j)})}{p(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(j)})}. \tag{2.5}$$

Posterior inferences at time $t+1$ can be based directly on these sampled values and weights, or we may resample according to the importance weights $\omega_{t+1}^{(j)}$ to obtain an equally weighted set of states representing a direct Monte Carlo approximation to the required posterior in equation (2.3).

2.3 Filtering for States and Parameters

In the general model with fixed parameters $\boldsymbol{\theta}$, extend the sample-based framework as follows. Standing at time t, we now have a combined sample

$$\{\boldsymbol{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)} : \ j = 1, \ldots, N\}$$

and associated weights

$$\{\omega_t^{(j)} : \ j = 1, \ldots, N\}$$

representing an importance sample approximation to the time t posterior $p(\boldsymbol{x}_t, \boldsymbol{\theta}|D_t)$ for both parameter and state. Note that the t suffix on the $\boldsymbol{\theta}$ samples here indicate that they are from the time t posterior, *not* that $\boldsymbol{\theta}$ is time-varying. Time evolves to $t+1$, we observe $\boldsymbol{y}_{t+1}$, and now want to generate a sample from $p(\boldsymbol{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_{t+1})$. Bayes' theorem gives this as

$$\begin{aligned} p(\boldsymbol{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_{t+1}) &\propto p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}, \boldsymbol{\theta}) p(\boldsymbol{x}_{t+1}, \boldsymbol{\theta}|\boldsymbol{D}_t) \\ &\propto p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}, \boldsymbol{\theta}) p(\boldsymbol{x}_{t+1}|\boldsymbol{\theta}, \boldsymbol{D}_t) p(\boldsymbol{\theta}|\boldsymbol{D}_t), \end{aligned} \tag{2.6}$$

where the form chosen in the last equation makes explicit the notion that the theoretical density function $p(\boldsymbol{\theta}|D_t)$ is an important ingredient in the update.

If $\boldsymbol{\theta}$ were known, equation (2.6) simplifies: $p(\boldsymbol{\theta}|D_t)$ is degenerate and we drop the known parameter from the conditioning statements. This leads to equation (2.3) and the auxiliary particle method applies for filtering on the state vector. Otherwise, it is now explicit that we have to deal with the problem of not knowing the form of the theoretical density function $p(\boldsymbol{\theta}|D_t)$ in order to obtain combined filtering on the parameter and the state. The next section reviews the two main historical approaches.

3 The Treatment of Model Parameters

3.1 Artificial Evolution of Parameters

In dealing with time-varying states, Gordon et al (1993) suggested an approach to reducing the sample degeneracy/attrition problem by adding small random disturbances (referred to as "roughening penalties") to state particles between time steps, in addition to any existing evolution noise contributions. In the literature since then, this idea has been extrapolated to fixed model parameters. One version of the idea adds small random perturbations to all the parameter particles under the posterior at each time point before evolving to the next. This specific method has an interpretation as arising from an extended model in which the model parameters are viewed as if they are, in fact, time-evolving – an "artificial evolution." That is, consider a different model in which $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\theta}_t$ at time t, and simply include $\boldsymbol{\theta}_t$ in an augmented state vector. Then add an independent, zero-mean normal increment to the parameter at each time. That is,

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t + \boldsymbol{\zeta}_{t+1} \\ \boldsymbol{\zeta}_{t+1} &\sim N(\mathbf{0}, \boldsymbol{W}_{\mathbf{t+1}}) \end{aligned} \tag{3.1}$$

for some specified variance matrix $\boldsymbol{W}_{t+1}$ and where $\boldsymbol{\theta}_t$ and $\boldsymbol{\zeta}_{t+1}$ are conditionally independent given D_t. With the model recast with the corresponding augmented state vector, the standard filtering methods for states alone, such as the auxiliary particle filter, now apply. The key motivating idea is that the artificial evolution provides the mechanism for generating new parameter values at each time step in the simulation analysis, so helping to address the sample attrition issue in reweighting methods that stay with the same sets of parameter points between time steps.

Among the various issues and drawbacks of this approach, the key one is simply that fixed model parameters are, well, fixed! Pretending that they are in fact time-varying implies an artificial "loss of information" between time points, resulting in posteriors that are, eventually, far too diffuse relative to the theoretical posteriors for the actual fixed parameters. To date there has been no resolution of this issue: if one adopts a model in which all parameters are subject to independent random shocks at each time point, the precision of resulting inferences is inevitably limited.

However, an inherent interpretation in terms of kernel smoothing of particles leads to a modification of this artificial evolution method in which the problem of information loss is avoided. First we discuss the basic form of kernel smoothing as introduced and developed in West (1993).

3.2 Kernel Smoothing of Parameters

Understanding the imperative to develop some method of smoothing for approximation of the required density $p(\boldsymbol{\theta}|D_t)$ in equation (2.6), West (1993b) developed kernel smoothing methods that provided the basis for rather effective adaptive importance sampling techniques. This represented extension to sequential analysis of basic mixture modelling ideas in West (1993a).

Standing at time t, suppose we have current posterior parameter samples $\boldsymbol{\theta}_t^{(j)}$ and weights $\omega_t^{(j)}$, $(j = 1, ..., N)$, providing a discrete Monte Carlo approximation to $p(\boldsymbol{\theta}|D_t)$. Again remember that the t suffix on $\boldsymbol{\theta}$ here indicates that the samples are from the time t posterior; $\boldsymbol{\theta}$ is not time-varying. Write $\overline{\boldsymbol{\theta}}_t$ and $\boldsymbol{V}_t$ for the Monte Carlo posterior mean and variance matrix of $p(\boldsymbol{\theta}|D_t)$, computed from the Monte Carlo sample $\boldsymbol{\theta}_t^{(j)}$ with weights $\omega_t^{(j)}$. The smooth kernel density form of West (1993a,b) is given by

$$p(\boldsymbol{\theta}|D_t) \approx \sum_{j=1}^{N} \omega_t^{(j)} N(\boldsymbol{\theta}|\boldsymbol{m}_t^{(j)}, h^2\boldsymbol{V}_t) \tag{3.2}$$

with the following defining components. First, $N(\cdot|\boldsymbol{m}, \boldsymbol{S})$ is a multivariate normal density mean $\boldsymbol{m}$ and variance matrix $\boldsymbol{S}$, so that the above density is a mixture of $N(\boldsymbol{\theta}|\boldsymbol{m}_t^{(j)}, h^2\boldsymbol{V}_t)$ distributions weighted by the sample weights $\omega_t^{(j)}$. Kernel rotation and scaling uses $\boldsymbol{V}_t$, the Monte Carlo posterior variance, and the overall scale of kernels is a function of the smoothing parameter $h > 0$. Standard density estimation methods suggest that h be chosen as a slowly decreasing function of N, so that kernel components are naturally more concentrated about their locations $\boldsymbol{m}_t^{(j)}$ for larger N. West (1993a,b) suggests taking slightly smaller values than the conventional kernel methods as a general rule. As we discuss below, our new work has led to a quite different perspective on this issue.

The kernel locations $\boldsymbol{m}_t^{(j)}$ are specified using a shrinkage rule introduced by West (1993a,b). Standard kernel methods would suggest $\boldsymbol{m}_t^{(j)} = \boldsymbol{\theta}_t^{(j)}$ so that kernels are located about existing sample values. However, this results in a kernel density function that is *over-dispersed* relative to the posterior sample, in the sense that the variance of the resulting mixture of normals is $(1 + h^2)\boldsymbol{V}_t$, always larger than $\boldsymbol{V}_t$. This is a most significant flaw in the sequential simulation; an over-dispersed approximation to $p(\boldsymbol{\theta}|D_t)$ will lead to an over-dispersed approximation to $p(\boldsymbol{\theta}|D_{t+1})$, and the consequent "loss of information" will build up as the operation is repeated at future times. To correct this, West (1993a,b) introduced the novel idea of shrinkage of kernel locations. Take

$$\boldsymbol{m}_t^{(j)} = a\boldsymbol{\theta}_t^{(j)} + (1 - a)\overline{\boldsymbol{\theta}}_t \tag{3.3}$$

where $a = \sqrt{1 - h^2}$. With these kernel locations, the resulting normal mixture retains the mean $\overline{\boldsymbol{\theta}}_t$ and now has the correct variance $\boldsymbol{V}_t$, hence the over-dispersion is trivially corrected.

3.3 Reinterpreting Artificial Parameter Evolutions

The undesirable "loss of information" implicit in equation (3.1) can be easily quantified. The Monte Carlo approximation $\{\boldsymbol{\theta}_t^{(j)}, \omega_t^{(j)}\}$ to $p(\boldsymbol{\theta}|D_t)$ has mean $\overline{\boldsymbol{\theta}}_t$ and variance matrix $\boldsymbol{V}_t$. Hence, in the evolution in equation (3.1) with the innovation $\boldsymbol{\zeta}_{t+1}$ independent of $\boldsymbol{\theta}_t$ as proposed, the implied prior $p(\boldsymbol{\theta}_{t+1}|D_t)$ has the correct mean $\overline{\boldsymbol{\theta}}_t$ but variance matrix $\boldsymbol{V}_t + \boldsymbol{W}_{t+1}$. The loss of information is explicitly represented by the component $\boldsymbol{W}_{t+1}$. Now, there is a close tie-in between this method and the kernel smoothing approach. To see this clearly, note that the Monte Carlo approximation to $p(\boldsymbol{\theta}_{t+1}|D_t)$ implied by equation (3.1) is also a kernel form, namely

$$p(\boldsymbol{\theta}_{t+1}|D_t) \approx \sum_{j=1}^{N} \omega_t^{(j)} N(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t^{(j)}, \boldsymbol{W}_{t+1}) \tag{3.4}$$

and this, as we have seen, is over-dispersed relative to the required or "target" variance $\boldsymbol{V}_t$.

It turns out that we can correct for this over-dispersion as follows. The key is to note that our kernel method is effective due to the use of location shrinkage. This shrinkage pushes samples $\boldsymbol{\theta}_t^{(j)}$ values towards their mean $\overline{\boldsymbol{\theta}}_t$ before adding a small degree of "noise" implied by the normal kernel. This suggests that the artificial evolution method should be modified by introducing correlations between $\boldsymbol{\theta}_t$ and the random shock $\boldsymbol{\zeta}_{t+1}$. Assuming a non-zero covariance matrix, note that the artificial evolution equation (3.1) implies

$$V(\boldsymbol{\theta}_{t+1}|D_t) = V(\boldsymbol{\theta}_t|D_t) + \boldsymbol{W}_{t+1} + 2C(\boldsymbol{\theta}_t, \boldsymbol{\zeta}_{t+1}|D_t).$$

To correct to "no information lost" implies that we set

$$V(\boldsymbol{\theta}_{t+1}|D_t) = V(\boldsymbol{\theta}_t|D_t) = \boldsymbol{V}_t,$$

which then implies

$$C(\boldsymbol{\theta}_t, \boldsymbol{\zeta}_{t+1}|D_t) = -\boldsymbol{W}_{t+1}/2.$$

Hence there must be a structure of negative correlations to remove the unwanted information loss effect. In the case of approximate joint normality of $(\boldsymbol{\theta}_t, \boldsymbol{\zeta}_{t+1}|D_t)$, this would then imply the conditional normal evolution in which

$$p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) = N(\boldsymbol{\theta}_{t+1}|\boldsymbol{A}_{t+1}\boldsymbol{\theta}_t + (\boldsymbol{I} - \boldsymbol{A}_{t+1})\overline{\boldsymbol{\theta}}_t, (\boldsymbol{I} - \boldsymbol{A}_{t+1}^2)\boldsymbol{V}_t) \tag{3.5}$$

where $\boldsymbol{A}_{t+1} = \boldsymbol{I} - \boldsymbol{W}_{t+1}\boldsymbol{V}_t^{-1}/2$.

The resulting Monte Carlo approximation to $p(\boldsymbol{\theta}_{t+1}|D_t)$ is then a generalised kernel form with complicated shrinkage patterns induced by the shrinkage matrix $\boldsymbol{A}_{t+1}$. We restrict here to the very special case in which the evolution variance matrix $\boldsymbol{W}_{t+1}$ is specified using a standard discount factor technique. Specifically, take

$$\boldsymbol{W}_{t+1} = \boldsymbol{V}_t(\frac{1}{\delta} - 1)$$

where δ is a discount factor in $(0, 1]$, typically around $0.95 - 0.99$. In this case, $\boldsymbol{A}_{t+1} = a\boldsymbol{I}$ with $a = (3\delta - 1)/2\delta$ and the conditional evolution density above reduces

$$p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) \sim N(\boldsymbol{\theta}_{t+1}|a\boldsymbol{\theta}_t + (1-a)\overline{\boldsymbol{\theta}}_t, h^2\boldsymbol{V}_t) \tag{3.6}$$

where $h^2 = 1 - a^2$, so that $h^2 = 1 - ((3\delta - 1)/2\delta)^2$, and we note that $a = \sqrt{1 - h^2}$. The resulting Monte Carlo approximation to $p(\boldsymbol{\theta}_{t+1}|D_t)$ is then precisely of the kernel form of equation (3.2), but now with a controlling smoothing parameter h specified directly via the discount factor.

We therefore have a version of the method of Gordon et al (1993) applied to parameters that connects directly with kernel smoothing with shrinkage. This justifies the basic idea of an artificial evolution for fixed model parameters in a modification that removes the problem of information loss over time.

Note also that the modified artificial evolution model of equation (3.6) may be adopted directly, without reference to the motivating discussion involving normal posteriors. This is clear from the following general result. Suppose $p(\boldsymbol{\theta}_t|D_t)$ has a finite mean $\overline{\boldsymbol{\theta}}_t$ and variance matrix $\boldsymbol{V}_t$, *whatever the global form of the distribution may be.* Suppose in addition that $\boldsymbol{\theta}_{t+1}$ is generated by the evolution model specified by equation (3.6). It is then easily seen that the mean and variance matrix of the implied marginal distribution $p(\boldsymbol{\theta}_{t+1}|D_t)$ are also $\overline{\boldsymbol{\theta}}_t$ and $\boldsymbol{V}_t$. Hence the connection with kernel smoothing with shrinkage, and the adjustment to fix the problem of information loss over time in artificial evolution approaches, is quite general. Note finally that the framework provides a direct method of specifying the scale of kernels via the single discount factor δ, as h (and a) are then simply determined. Generally, a higher discount factor – around 0.99 – will be relevant.

4 A General Algorithm

Return now to the general filtering problem, that of sampling the posterior in equation (2.6). We have available the Monte Carlo sample $(\boldsymbol{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)})$ and weights $\omega_t^{(j)}$ $(j = 1, \ldots, N)$, representing the joint posterior $p(\boldsymbol{x}_t, \boldsymbol{\theta}|\boldsymbol{D}_t)$. Again, the suffix t on the parameter samples indicates the time t posterior, not time-variation. We adopt the kernel form of equation (3.2) as the marginal density for the parameter, following the earlier discussion. With the equivalent interpretation of this as arising from an artificial evolution with correlation structure, as just discussed, we can now apply an extended version of the auxiliary particle filter algorithm, incorporating the parameter with the state. The resulting general algorithm runs as follows.

1. For each $j = 1, ..., N$, identify the prior point estimates of $(\boldsymbol{x}_t, \boldsymbol{\theta})$ given by $(\boldsymbol{\mu}_{t+1}^{(j)}, \boldsymbol{m}_t^{(j)})$ where

$$\boldsymbol{\mu}_{t+1}^{(j)} = E(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t^{(j)}, \boldsymbol{\theta}_t^{(j)}).$$

may be computed from the state evolution density and $\boldsymbol{m}_t^{(j)} = a\boldsymbol{\theta}_t^{(j)} + (1-a)\overline{\boldsymbol{\theta}}_t$ is the j^{th} kernel location from equation (3.3).

2. Sample an auxiliary integer variable from the set $\{1, ..., N\}$ with probabilities proportional to

$$g_{t+1}^{(j)} \propto \omega_t^{(j)} p(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(j)}, \boldsymbol{m}_t^{(j)});$$

call the sampled index k.

3. Sample a new parameter vector $\boldsymbol{\theta}_{t+1}^{(k)}$ from the k^{th} normal component of the kernel density, namely

$$\boldsymbol{\theta}_{t+1}^{(k)} \sim N(\cdot|\boldsymbol{m}_t^{(k)}, h^2\boldsymbol{V}_t).$$

4. Sample a value of the current state vector $\boldsymbol{x}_{t+1}^{(k)}$ from the system equation

$$p(\cdot|\boldsymbol{x}_t^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)}).$$

5. Evaluate the corresponding weight

$$\omega_{t+1}^{(k)} \propto \frac{p(\boldsymbol{y}_{t+1}|\boldsymbol{x}_{t+1}^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)})}{p(\boldsymbol{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(k)}, \boldsymbol{m}_t^{(k)})}.$$

6. Repeat step (2)-(5) a large number of times to produce a final posterior approximation $(\boldsymbol{x}_{t+1}^{(k)}, \boldsymbol{\theta}_{t+1}^{(k)})$ with weights $\omega_{t+1}^{(k)}$, as required.

Note that the Monte Carlo sample size N can be different at each time point, if required. Also, we might be interested in over-sampling a rather larger set of values and then resampling according to the weights above in order to produce an equally weighted final sample. Further, it is generally appropriate to operate with parameters that are real-valued, when using the normal kernel method. Hence we routinely deal with log variances rather than variances, logit transforms of parameters restricted to a finite range – such as the autoregressive parameter in the following AR(1) example and the later dynamic factor model – and so forth.

Example

As a simple example in which the sequential updating is available in closed form for comparison, consider the AR(1) model in which $\boldsymbol{y}_t = x_t$, a scalar, with $x_{t+1} \sim N(x_t\phi, 1)$. Here $\boldsymbol{\theta} = \phi$, a single parameter, and there is no unobserved state variable. Hence the focus is exclusively on the efficacy of learning the parameter (steps 1 and 4 of the general algorithm above are vacuous).

A realisation of length 897 was generated from this AR(1) model at $\phi = 0.8$. The sequential analysis was then performed over times $t = 1, \ldots, 897$, and the posterior approximation at $t = 897$ compared to the exact (normal) posterior for ϕ. The simulation-based analysis used $N = 5000$ sample points throughout. Figure 1 graphs the time trajectories of the posterior sample quantiles (2.5%, 25%, 50%, 75%, 97.5%) together with the exact posterior quantiles. Agreement is remarkable across the entire time period, with very little evidence of "build up" of approximation error. Table 1 provide a numerical comparison of exact and approximate quantiles for $p(\phi|D_{897})$ which further illustrates the accuracy. By comparison, a direct use of the artificial evolution method leads to a gradual decay of approximation accuracy due to the loss of information.

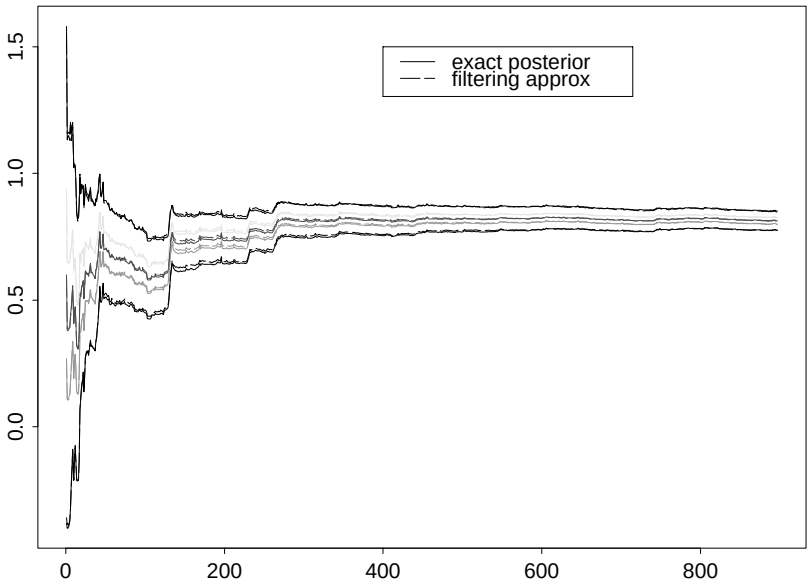

Figure 1. Time trajectories of posterior quantiles (2.5%, 25%, 50%, 75%, 97.5%) of the posteriors for the AR(1) parameter ϕ.

	0.025	0.25	0.50	0.75	0.975
exact:	0.7755	0.8005	0.8136	0.8267	0.8517
approx:	0.7758	0.7993	0.8119	0.8242	0.8482

Table 1. Posterior quantiles from the posteriors for the AR(1) parameter ϕ at $t = 897$.

5 Factor Stochastic Volatility Modelling

In studies of dynamic latent factor models with multivariate stochastic volatility components, recent Bayesian work has developed both MCMC methods and aspects of sequential analysis using versions of auxiliary particle filtering for states (Aguilar and West 2000, Pitt and Shephard 1999a). In these models, the state variables are latent volatilities of both common factor processes and of residual/idiosyncratic random terms specific to observed time series. In the application in exchange rate modelling, forecasting and portfolio analysis, Aguilar and West (1998) use MCMC methods to fit these complicated dynamic models to historical data, and then perform sequential particle filtering over a long stretch of further data that provides the context for sequential forecasting and portfolio construction. In that example, these authors fix a full set of constant model parameters at estimated values taken as the means of posterior distributions based on the MCMC analysis of the initial (and very long) data stretch. The results are very positive from the financial time series modelling viewpoint. For many practical purposes, an extension of that approach that involves periodic reanalysis of some recent historical data using full MCMC methods, followed by sequential analysis using auxiliary particle filtering on just the time-varying states with model parameters fixed at most recently estimated values, is quite satisfactory. However, from the viewpoint of the use of sequential simulation technology in more interesting, and complicated models, this setting provides a very nice and somewhat challenging test-bed, especially when considering multivariate time series on moderate dimensions. Hence our interest in exploring the general sequential algorithm of the previous section in this context.

We adopt the context and notation of Aguilar and West (1998), noting the very similar developments in Pitt and Shephard (1999a) and that our models are based on those of the earlier authors (Kim, Shephard and Chib 1998). Begin with a $q-$variate time series of observations, in $\boldsymbol{y}_t$, $(t = 1, 2, \ldots)$. In our example, this is a vector of observed daily exchange rates of a set of $q = 6$ national currencies relative to the US dollar. The dynamic latent factor model with multivariate stochastic volatility components is defined and structured as follows.

At each time t, we have

$$\boldsymbol{y}_t = \boldsymbol{\alpha}_t + \boldsymbol{X}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t \tag{5.1}$$

with the following ingredients.

- $\boldsymbol{y}_t$ is the q-vector of observation and $\boldsymbol{\alpha}_t$ is a q-vector representing a local level of the series.
- $\boldsymbol{X}$ is a $q \times k$ matrix called the *factor loadings matrix*.

- $\boldsymbol{f}_t$ is a $k-$vector which represents the vector of *latent factors* at time t; the $\boldsymbol{f}_t$ are assumed to be conditionally independent over time and distributed as $N(\boldsymbol{f}_t|\boldsymbol{0}, \boldsymbol{H}_t)$ where $\boldsymbol{H}_t = \text{diag}(h_{t1}, \ldots, h_{tk})$ is the diagonal matrix of instantaneous factor variances.

- $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{\epsilon}_t|\boldsymbol{0}, \boldsymbol{\Psi})$ are idiosyncratic noise terms, assumed to be conditionally independent over time and with a diagonal variance matrix $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_k)$. The elements of $\boldsymbol{\Psi}$ are called the *idiosyncratic noise variances* of the series. We note that Aguilar and West (1998) use an extension of this model that, as in Pitt and Shephard (1999a), has time-varying idiosyncratic noise variances, but we do not consider that here.

- $\boldsymbol{\epsilon}_t$ and $\boldsymbol{f}_s$ are mutually independent for all t, s.

Following earlier authors (Geweke and Zhou 1996), Aguilar and West (1998) adopt a factor loading matrix of the form

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ x_{2,1} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ x_{k,1} & x_{k,2} & x_{k,3} & \cdots & 1 \\ x_{k+1,1} & x_{k+1,2} & x_{k+1,3} & \cdots & x_{k+1,k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{q,1} & x_{q,2} & x_{q,3} & \cdots & x_{q,k} \end{pmatrix}. \tag{5.2}$$

The reduced number of parameters in $\boldsymbol{X}$ ensures mathematical identification of the model and the lower-triangular form provides a nominal identification of the factors: the first series is driven by the first factor alone, the second series is driven by the first two factors, and so forth.

Stochastic volatility structures are defined for the sequences of conditional variances of the factors. For each $i = 1, \ldots, k$, define $\lambda_{ti} = \log(h_{ti})$, and write $\boldsymbol{\lambda}_t = (\lambda_{t1}, \ldots, \lambda_{tk})$. The set of log factor variances $\{\boldsymbol{\lambda}_t\}$ is modelled as a vector autoregression of order one, VAR(1), to capture correlations in fluctuations in volatility levels. Specifically,

$$\boldsymbol{\lambda}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}(\boldsymbol{\lambda}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{\gamma}_t \tag{5.3}$$

with the following ingredients: $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ is the underlying stationary volatility level, $\boldsymbol{\Phi} = \text{diag}(\phi_1, \ldots, \phi_k)$ is a diagonal matrix with individual AR parameters ϕ_i for factor volatility process λ_{ti}, and the innovations vectors $\boldsymbol{\gamma}_t$ are conditionally independent and normal,

$$\boldsymbol{\gamma}_t \sim N(\boldsymbol{\gamma}_t|\boldsymbol{0}, \boldsymbol{U}) \tag{5.4}$$

for some innovations variance matrix $\boldsymbol{U}$. This model differs from that of Pitt and Shephard (1999a) in several respects, an important one being that we

allow non-zero off-diagonal entries in $\boldsymbol{U}$ to estimate dependencies in changes in volatility patterns across the factors. This turns out to be empirically supported and practically relevant in short-term exchange rate modelling. We note that Aguilar and West (1998) also develop stochastic volatility model components for the variances $\boldsymbol{\Psi}$ of the idiosyncratic errors, but we do not explore that here.

We analyse the one-day-ahead returns on exchange rates over a period of several years in the 1990s, as in Aguilar and West (1998). Taking s_{ti} as the spot rate in US dollars for currency i on day t, the returns are simply $y_{ti} = s_{ti}/s_{t-1,i} - 1$ for currency $i = 1, \ldots, q = 6$. The currencies are, in order, the Deutschmark/Mark (DEM), Japanese Yen (JPY), Canadian Dollar (CAD), French Franc (FRF), British Pound (GBP) and Spanish Peseta (ESP). Here we explore analysis of the returns over the period 12/1/92 to 8/9/96, a total of 964 observations. We adopt the model as structured above, and take an assumedly fixed return level $\boldsymbol{\alpha} = \boldsymbol{\alpha}$.

We first performed intensive Bayesian analysis of the first 914 observations using the MCMC simulation approach of Aguilar and West (1998). At $t = 914$, we then have a full sample from the actual posterior, based on data up to that time point, for all past latent factors, their volatilities, and all fixed model parameters. In terms of proceeding ahead sequentially, we identify the relevant state variables and parameters as follows. First note that we can reduce the model equation (5.1) by integrating out the latent factors to give the conditional observation distribution

$$\boldsymbol{y}_t \sim N(\boldsymbol{y}_t | \boldsymbol{0}, \boldsymbol{X}\boldsymbol{H}_t\boldsymbol{X}' + \boldsymbol{\Psi}). \tag{5.5}$$

Now introduce the definitions of the state variable

$$\boldsymbol{x}_t \equiv \boldsymbol{H}_t$$

at time t, and the fixed model parameters

$$\boldsymbol{\theta} = \{\boldsymbol{X}, \boldsymbol{\Psi}, \boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{U}\}.$$

In our example, the state variable $\boldsymbol{x}_t$ is $3-$dimensional, and the parameter $\boldsymbol{\theta}$ is $36-$dimensional. As we discuss below, the sequential component of the study reported here treats $\boldsymbol{U}$ as fixed at a value based on the MCMC analysis of the first 914 observations, so that $\boldsymbol{\theta}$ reduces to 30 free model parameters, and the posterior at each time point is in 33 dimensions. For reference, the estimate of $\boldsymbol{U}$ is

$$E(\boldsymbol{U}|D_{914}) = \begin{pmatrix} 0.0171 & 0.0027 & 0.0009 \\ 0.0027 & 0.0194 & 0.0013 \\ 0.0009 & 0.0013 & 0.0174 \end{pmatrix}$$

based on the initial MCMC analysis over $t = 1, \ldots, 914$. The posterior standard deviations of elements of $\boldsymbol{U}$ at $t = 914$ are all of the order of 0.001-0.003,

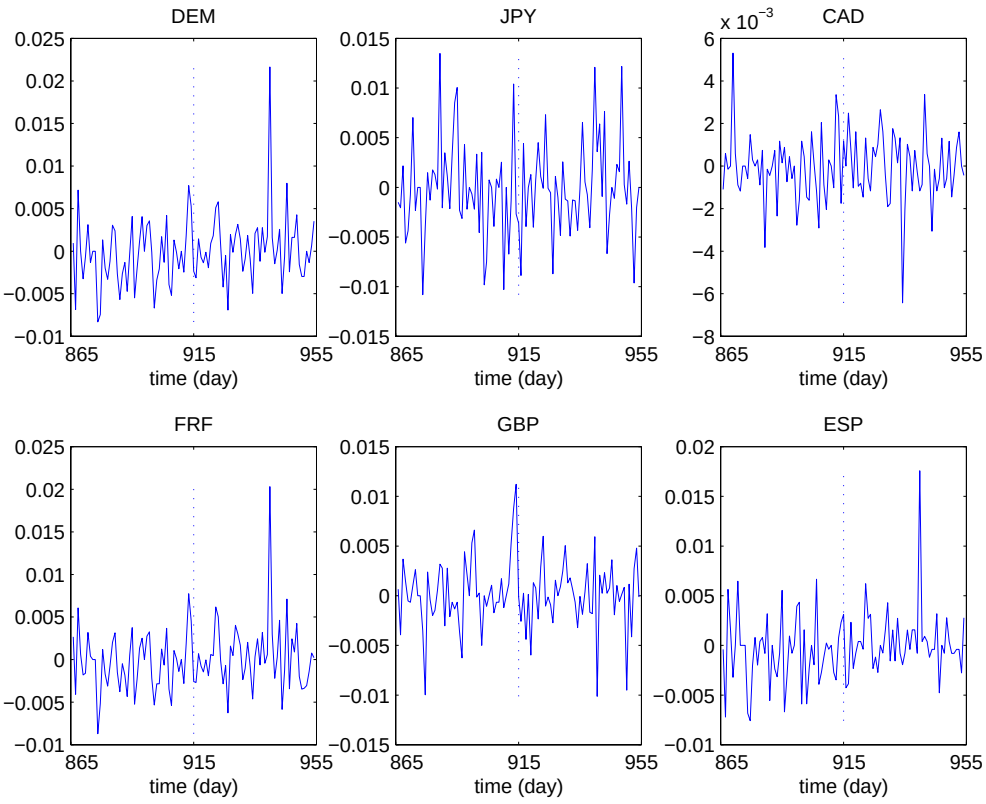

Figure 2. Exchange rate time series

so there is a fair degree of uncertainty about $\boldsymbol{U}$ that is being ignored in the sequential analysis and comparison.

To connect with the general dynamic model framework, note that we now have the observation equation (2.1) defined by the model equation (5.5), and the evolution equation (2.2) given implicitly by the stochastic volatility model equations (5.3) and (5.4). We can now apply the general sequential filtering algorithm, and do so starting at time $t = 914$ with the full posterior $p(\boldsymbol{x}_{914}, \boldsymbol{\theta}|D_{914})$ available as a large Monte Carlo sample based on the MCMC analysis of all the data up to that time. It is relevant to note that the context here – with an informed prior $p(\boldsymbol{x}_{914}, \boldsymbol{\theta}|D_{914})$ based on past data, is precisely that facing practical analysts in many fields in which further analysis, at least over short stretches of data, is required to be sequential. As noted above, we make one change: for reasons discussed below we fix the VAR(1) innovations variance matrix $\boldsymbol{U}$ at the estimate $E(\boldsymbol{U}|D_{914})$ and so the parameter $\boldsymbol{\theta}$ is reduced by removal of $\boldsymbol{U}$. Hence the particle filtering applies to the 3 state variables at each time point and the 30 model parameters. We then proceed to analyse further data, sequentially, over $t = 915, 916, \ldots$. Figure 2 displays a stretch of the data running from $t = 864$ to $t = 964$ with $t = 914$ marked. Our sequential filtering methods produces Monte Carlo approximations to

each $p(\boldsymbol{x}_t, \boldsymbol{\theta}|D_t)$ over $t = 915, 916, \ldots$. Throughout the analysis, the Monte Carlo sample size is fixed at $N = 9000$ at each step. The kernel shrinkage and shapes are defined via the discount factor $\delta = 0.99$ which implies $a = 0.995$ and $h = 0.1$. A final technical point to note is that we operate with the kernel method on parameters transformed so that normal kernels are appropriate; thus each of the μ_j and ϕ_j parameters is transformed to the logit scale, and the variance parameters ψ_j are logged (this follows West 1993a,b).

Our experiences in this study mirror those of using the straight auxiliary particle filtering method when the parameters are assumed fixed (Aguilar and West 2000). That is, filtering on the volatilities is a more-or-less standard problem, and the state variable is in only 3 dimensions so performance is expected to be excellent. The questions of accuracy and performance in the extended context with a larger number of parameters are now much more interesting, however, due to the difficulties inherent in dealing with discrete samples in higher dimensional parameter spaces. Inevitably, the accuracy of approximation is degraded relative to simple filtering on two or three time-evolving states. One way to define "performance" here is via comparison of the sequentially computed Monte Carlo approximations to posteriors with those based on a full MCMC analysis that refits the entire data set up to specified time points of interest. Our discussion here focuses specifically on this aspect in connection with inferences on the fixed model parameters. For a chosen set of times during the period of 50 observations between $t = 914$ and 964 we re-ran the full MCMC analysis of the factor model based on all the data up to that time point, and explored comparisons with the sequential filtering-based approximations in which we begin filtering at $t = 914$. At any time t the posterior from the MCMC analysis plays the role of the "true" posterior, or at least the "gold standard" by which to assess the performance of the filtering algorithm. Some relevant summaries appear in Figures 3 to 12 inclusive. The first set, Figures 3–7, display summaries of the univariate marginal posterior distributions at $t = 924$. We refer to this as the 10-step analysis, as the sequential filter is run for just 10 steps from the starting position at $t = 914$. For each of the 30 fixed parameters, we display quantile plots comparing quantiles of the approximate posteriors from the MCMC and sequential analyses. The graphs indicate (with crosses) the posterior quantiles at 1, 5, 25, 50, 75, 95 and 99% of the posteriors, graphing the filtering-based quantiles versus those from the MCMC. The $y = x$ line is also drawn. From these graphs, it is evident that posterior margins are in excellent agreement (we could have added approximate intervals to these plots, based on methods of Bayesian density estimation, to represent uncertainty about the estimated quantile functions; for the large sample sizes of 9000 here, such intervals are extremely narrow except in the extreme tails, and just obscure the plots.) The poster margins computed by the auxiliary particle filtering (APF) and MCMC analyses are the same for all practical purposes. Only in the very extreme upper tail of two of the VAR model parameters – μ_1 and the logit

of ϕ_1 – are there any deviations at all, and here the APF posterior is very slightly heavier tailed than that from the MCMC, but the differences are hardly worth a mention.

The remaining graphs, Figures 8–12, display similar quantile plots comparing the APF and MCMC posteriors $t = 964$. This provides a similar comparison but now for a 50-step analysis, the sequential filter running for 50 time points from the starting position at $t = 914$. Again we see a high degree of concordance in general, although the longer filtering period has introduced some discrepancies between the marginal posteriors, especially in the extreme tails of several of the margins. Some of the bigger apparent differences appear in the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\mu}$ of the VAR volatility model component, indicated in Figures 11 and 12. Also noteworthy is the fact that this period of 50 observations includes a point at around $t = 940$ where the series exhibits a real outlier, peaking markedly in the DEM, FRF, ESP and CAN series. Such events challenge sequential methods of any kind, and may play a role here in inducing small additional inaccuracies in the APF approximations by skewing the distribution of posterior weights at that time point. We do have ranges of relevant methods for model monitoring and adaptation to handle such events (West 1986, West and Harrison 1986, 1989 and 1997) though such methods are not applied in this study.

One additional aspect of the analysis worth noting is that the distributions of the sets of sequentially updated weights $\omega_t^{(j)}$ remain very well-behaved across the 50 update points. The shape is smooth and unimodal near the norm of $1/N = 1/9000$, with few weights deviating really far at all. Even at the outlier point the maximum weight is only is 0.004, fewer than 200 of the 9000 weights are less than 0.1/9000, and only 16 exceed 10/9000. All in all, we can view the analysis as indicating the utility of the filtering approach even over rather longer time intervals.

As earlier mentioned, the sequential simulation analysis fixes the volatility innovations variance matrix $\boldsymbol{U}$ at its prior mean $E(\boldsymbol{U}|D_{914})$. This is because we have no easy way of incorporating a structured set of parameters such as $\boldsymbol{U}$ in the kernel framework – normal distributions do not apply to symmetric positive definite matrices of parameters. It may be that fixing this set of parameters induces some inaccuracies in the filtering analysis compared to the MCMC analysis. With this in mind, we should expect to see some differences between the APF posterior and the MCMC, and these differences can be expected to be most marked in the margins for the volatility model parameters $\boldsymbol{\mu}$ and, most particularly, $\boldsymbol{\phi}$. The largest differences in the 50-step analysis do indeed relate to $\boldsymbol{\phi}$, suggesting that some of the differences generally may indeed be due to the lack of proper accounting for the uncertainties about $\boldsymbol{U}$. Looking ahead, it is of interest to anticipate developments of kernel methods that allow for such structuring – perhaps using normal kernels with elaborate reparametrisations, or perhaps with a combination of normal and non-normal kernels – though for the moment we have no way of doing this.

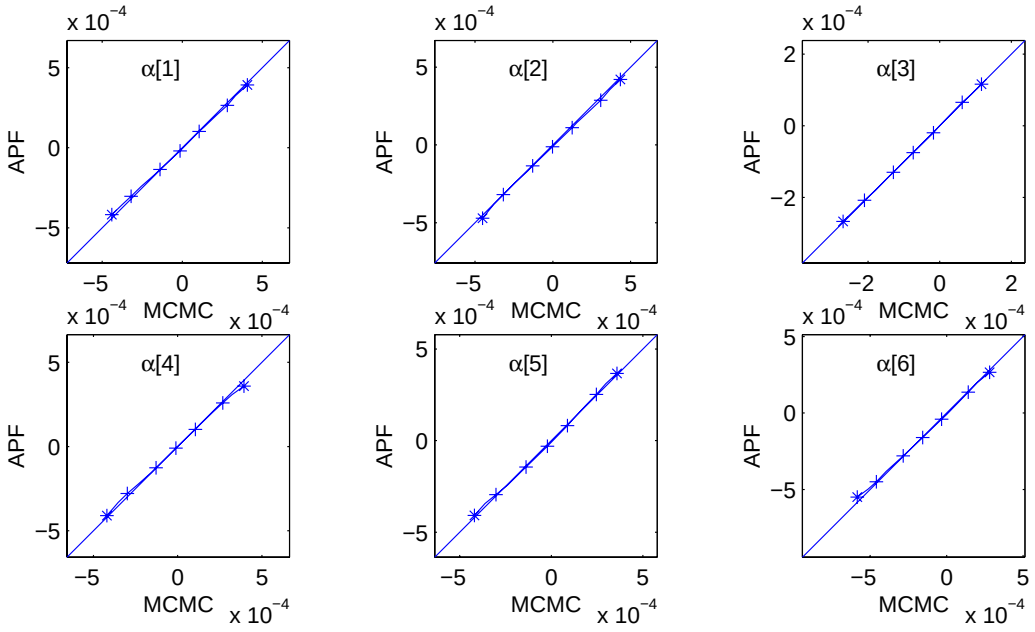

Figure 3. Q-Q plots of posterior samples of the α_j parameters in the 10-step analysis

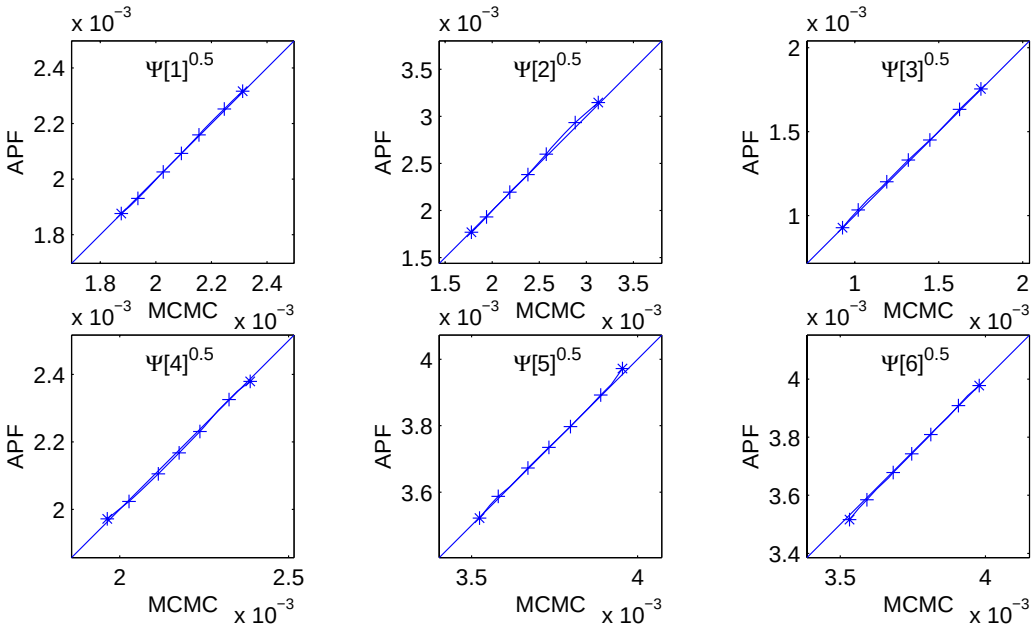

Figure 4. Q-Q plots of posterior samples of the $\sqrt{\psi_j}$ parameters in the 10-step analysis

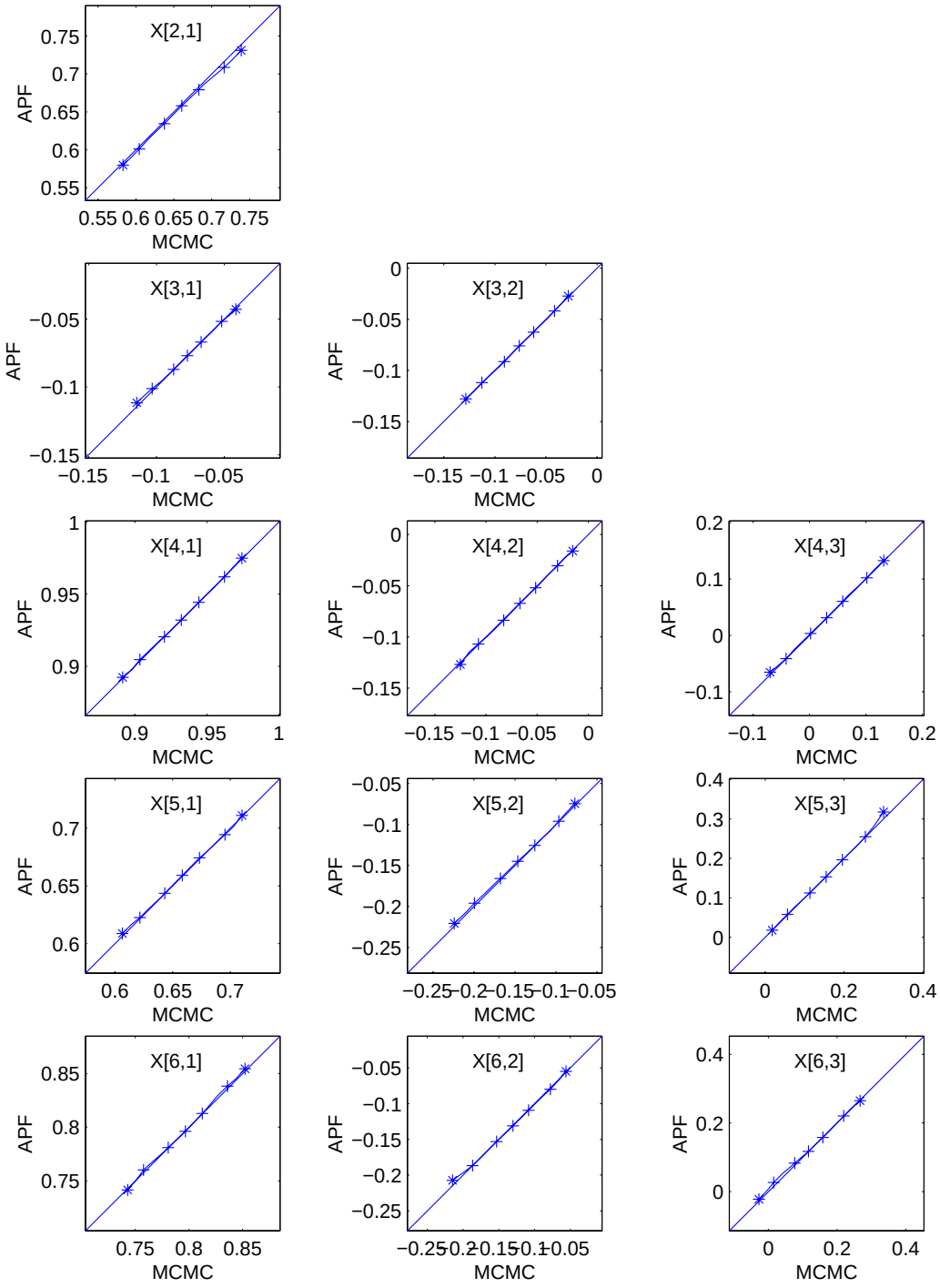

Figure 5. Q-Q plots of posterior samples of the X_{ij} parameters in the 10-step analysis

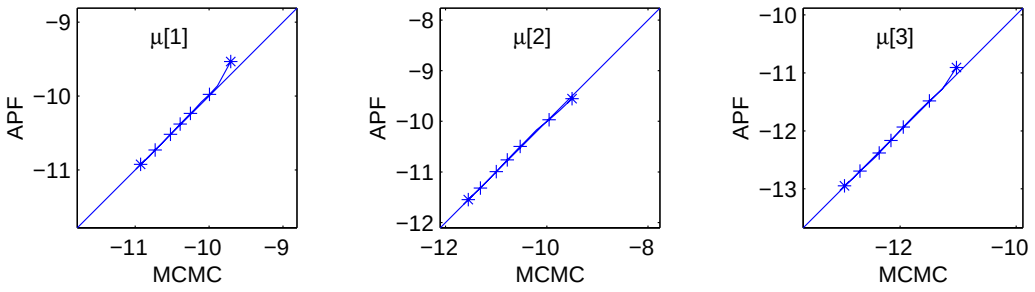

Figure 6. Q-Q plots of posterior samples of the μ_j parameters in the 10-step analysis

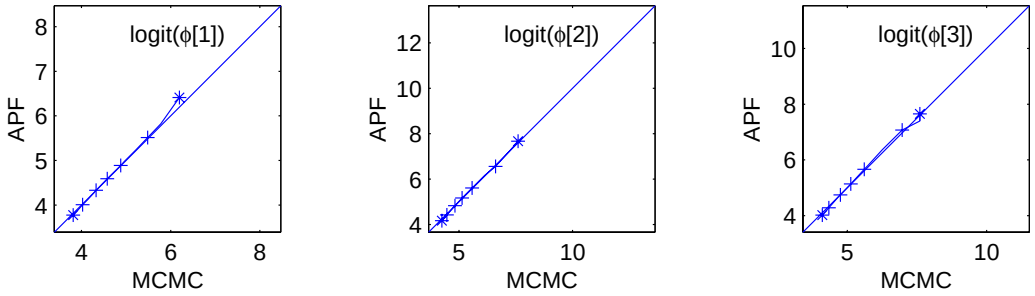

Figure 7. Q-Q plots of posterior samples of the logits of the ϕ_j parameters in the 10-step analysis

6 Discussion and Future Directions

In the moderate dimensional model above, the analysis certainly indicates the feasibility of sequential simulation-based filtering using the extended auxiliary particle filtering algorithm that incorporates several parameters in addition to state variables. Performance relative to the (almost) equivalent MCMC analysis is excellent; for most practical purposes, the results are in good agreement with the MCMC results even in the 50-step analysis where some minor differences in tail behaviour are noted. We have indicated some possible reasons for these differences that are not related to the specific algorithm nor the sequential context. If we ignore those issues and assume that all differences arise due to the inaccuracies inherent in sequential particle filtering, it is clear that there should be room for improvement. Before discussing some ideas and suggestions for improvements, we want to stress the relevance of context and goals. Sequential filtering inherently induces approximation errors that may tend to build up over time. In applied work, such as in using dynamic factor models in financial analysis, this must be accounted for and corrected. In existing application of factor models with collaborators in the banking industry, the sequential filtering methods are used over only short

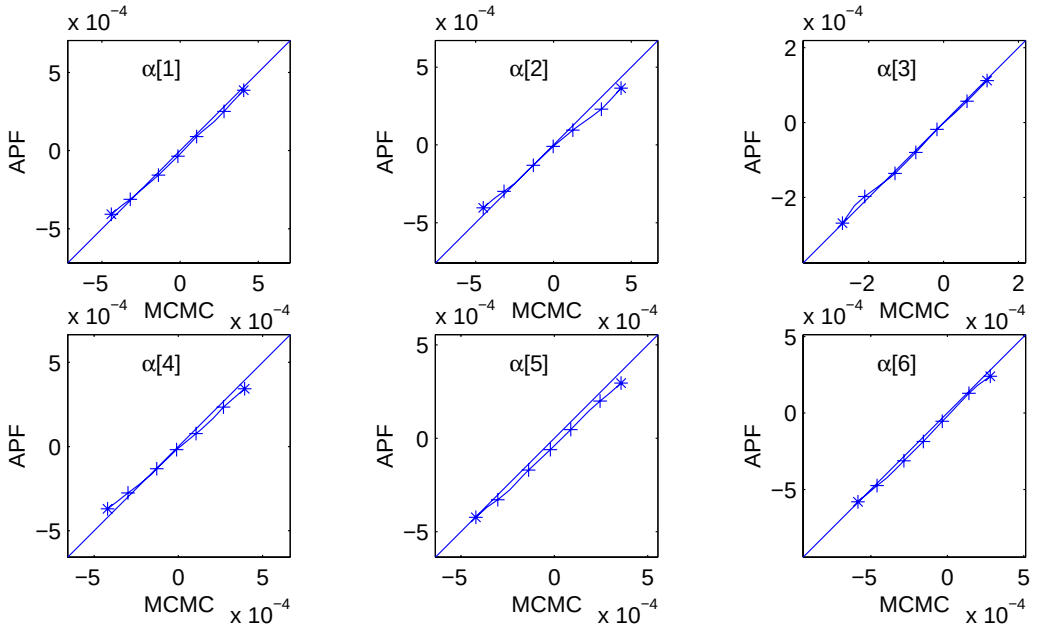

Figure 8. Q-Q plots of posterior samples of the α_j parameters in the 50-step analysis

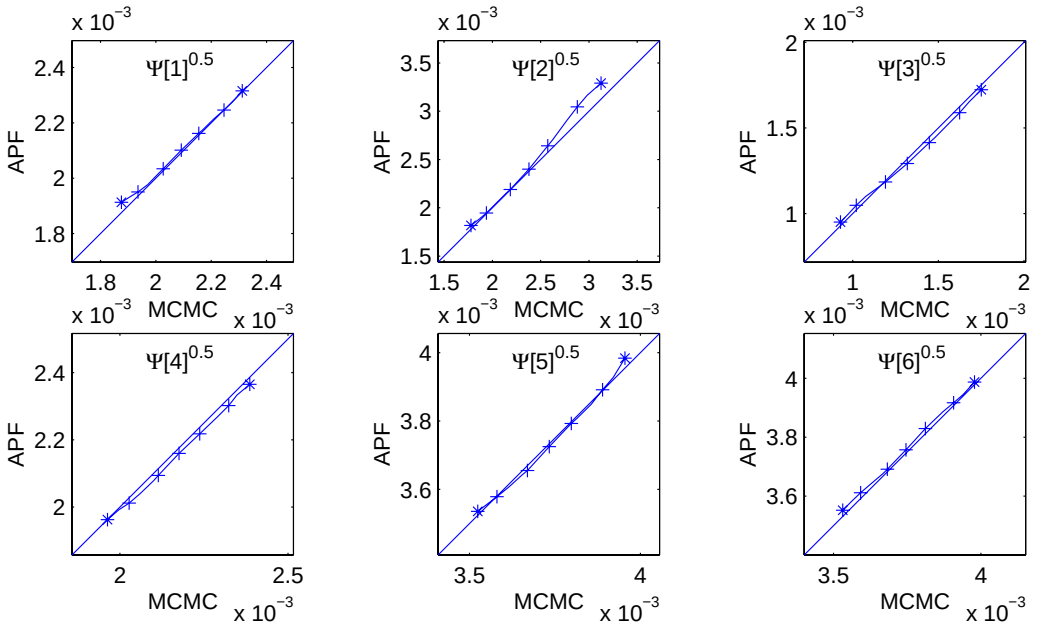

Figure 9. Q-Q plots of posterior samples of the $\sqrt{\psi_j}$ parameters in the 50-step analysis

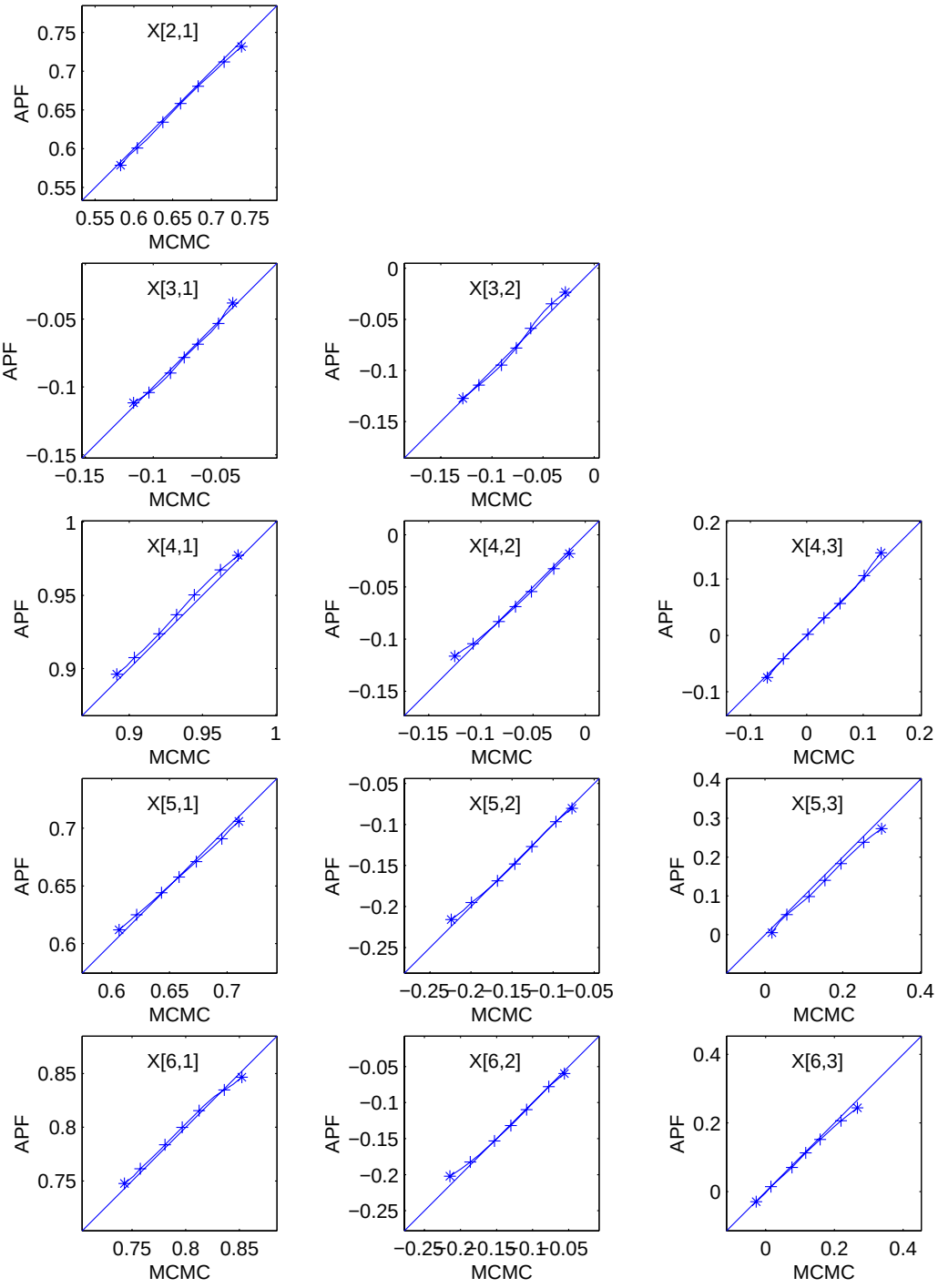

Figure 10. Q-Q plots of posterior samples of the X_{ij} parameters in the 50-step analysis

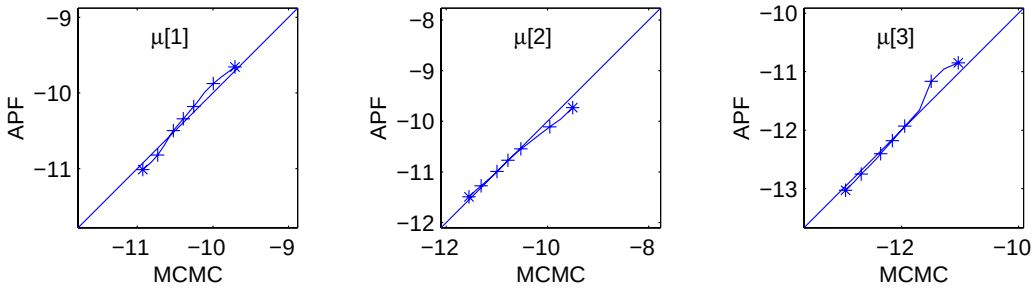

Figure 11. Q-Q plots of posterior samples of the μ_j parameters in the 50-step analysis

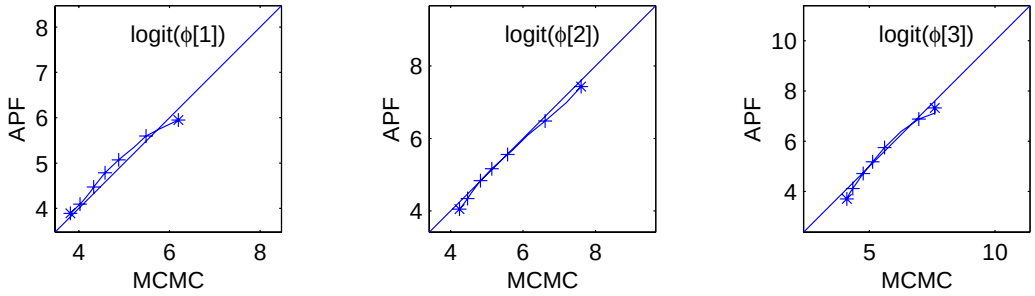

Figure 12. Q-Q plots of posterior samples of the logits of the ϕ_j parameters in the 50-step analysis

time scales – the 5 days of a working week with daily time series such as the exchange rate returns here. This is coupled with periodic updating based on a full MCMC analysis of a longer historical stretch of data (i.e., MCMC at the weekend based on the last several months of data). The horizon of 10 days in the example is therefore very relevant, whereas the 50 day horizon is very long and perhaps unrealistic. In this context, the differences between the filtering-based and MCMC-based posterior quantile functions at lower time steps are quite negligible relative to those at the longer time step. This experience and perspective is consistent with our long-held view that sequential simulation-based filtering methods must always be combined with some form of periodic re-calibration based on off-line analysis performed with much more computational time available than the filtering methods are designed to accommodate.

Some final comments relate to possible extensions of the filtering algorithm that may improve posterior approximations. Questions of accuracy and adequacy arise in connection with the approximation of (typical) posteriors that exhibit varying patterns of dependencies among parameters in different regions of the parameter space, and also varying patterns of tail-weight. Discrete, sample-based approximations inevitably suffer problems of generating

points far enough into the tails of fatter-tailed posteriors, especially in higher dimensions. It is sometimes helpful to use fatter-tailed kernels, such as T kernels (West 1993a, West 1993b) but this does not often help much and goes no way to addressing the real need for more sensitive analytic approximation of *local* structure in the posterior; the kernel mixture of equation (3.2) is *global* in that the mixture components are based on the same "global" shrinkage center $\overline{\boldsymbol{\theta}}_t$ and each have the same scales and shapes as determined by $h^2\boldsymbol{V}_t$. Very large numbers of such kernels are needed to truly adequately approximate posteriors that may evidence tails of differing weight in different dimensions (and fatter than normal tails), highly non-linear relationships among the parameters and hence varying patterns of "local" scale and shape as we move around in $\boldsymbol{\theta}$ space. We need to complement this suggestion with modifications that allow "differential shrinkage centers." West (1993a,b) discussed some of these issues, with suggestions about kernel methods with kernel-specific variance matrices in components in particular, and this idea was developed and implemented in certain non-sequential contexts in Givens and Raftery (1996). Development for implementation in sequential contexts remains an important research challenge.

A simple example helps to highlight these issues and underscores some suggestions for algorithmic extensions that follows. Consider a bimodal prior $p(\boldsymbol{\theta}|D_t)$ in which one mode has the shape of a unit normal distribution and the other that of a normal but with a much larger variance. In using a kernel approximation based on a prior sample, we would expect to do very much better using bimodal mixture in which sample points "near" one mode are shrunk towards that mode, and with kernel scalings that are higher for points "near" the second mode. The existing global kernel method uses global shrinkage to match the first two moments, but loses accuracy in less regular situations as this example indicates.

A specific research direction that reflects these considerations completes our discussion. To begin, consider the existing framework and recall that *any* density function $p(\boldsymbol{\theta}_t|D_t)$ may be arbitrarily well approximated by a mixture of normal distributions. Suppose therefore, for theoretical discussion, that the density has exactly such a form, namely

$$p(\boldsymbol{\theta}_t|D_t) = \sum_{r=1}^{R} q_r N(\boldsymbol{\theta}_t|\boldsymbol{b}_r, \boldsymbol{B}_r) \tag{6.1}$$

for some parameters R and $\{q_r, \boldsymbol{b}_r, \boldsymbol{B}_r : r = 1, \ldots, R\}$ (these will all depend on t though this is not made explicit in the notation, for clarity). In this case, the mean $\overline{\boldsymbol{\theta}}_t$ and variance matrix $\boldsymbol{V}_t$ are given by $\overline{\boldsymbol{\theta}}_t = \sum_{r=1}^{R} q_r \boldsymbol{b}_r$ and $\boldsymbol{V}_t = \sum_{r=1}^{R} q_r\{\boldsymbol{B}_r + (\boldsymbol{b}_r - \overline{\boldsymbol{\theta}}_t)(\boldsymbol{b}_r - \overline{\boldsymbol{\theta}}_t)'\}$. Suppose also, again, that $\boldsymbol{\theta}_{t+1}$ is generated by the evolution model specified by equation (3.6). It then easily

follows that the implied marginal density $p(\boldsymbol{\theta}_{t+1}|D_t)$ is of the form

$$p(\boldsymbol{\theta}_{t+1}|D_t) = \sum_{r=1}^{R} q_k N(\boldsymbol{\theta}_{t+1}|a\boldsymbol{b}_r + (1-a)\overline{\boldsymbol{\theta}}_t, a^2\boldsymbol{B}_r + (1-a^2)\boldsymbol{V}_t). \qquad (6.2)$$

Now, in general, this is not the same as the density of $(\boldsymbol{\theta}_t|D_t)$, though the mean and variance matrix match as mentioned above. In practice, a will be quite close to 1 so that the two distributions will be close, but not precisely the same in general. The exception is the case of a normal $p(\boldsymbol{\theta}_t|D_t)$, i.e., the case $R = 1$, when both $p(\boldsymbol{\theta}_t|D_t)$ and $p(\boldsymbol{\theta}_{t+1}|D_t)$ are $N(\cdot|\overline{\boldsymbol{\theta}}_t, \boldsymbol{V}_t)$. Otherwise, in the case of quite non-normal priors, the component means $\boldsymbol{b}_r$ will be quite separated, the local variance matrices $\boldsymbol{B}_r$ quite different in scale and structure. Hence the location and scale/shape shrinkage effects in the components of the resulting mixture (6.2) tend to obscure the differences by the implied shrinking/averaging.

This discussion, linking back to the important use of normality of $\boldsymbol{\theta}_t$ in the theoretical tie-up between artificial evolution methods and kernel methods in Section 3.3, suggests the following development. Suppose that the distribution $p(\boldsymbol{\theta}_t|D_t)$ is indeed exactly of the form of equation (6.2). To focus on "local" structure in this distribution, introduce the component indicator variable r_t such that $r_t = r$ with probability q_r $(r = 1, \ldots, R)$. Then $(\boldsymbol{\theta}_t|r_t = r, D_t) \sim N(\cdot|\boldsymbol{b}_r, \boldsymbol{B}_r)$. At this point we can apply the same line of reasoning about an artificial evolution to smooth a set of $\boldsymbol{\theta}_t$ samples, but now explicitly including the indicator r_t provides a focus on the *local* structure. This suggests the modification of the key evolution equation (3.6) to the local form

$$p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t, r_t = r, D_t) \sim N(\boldsymbol{\theta}_{t+1}|a\boldsymbol{\theta}_t + (1-a)\boldsymbol{b}_r, h^2\boldsymbol{B}_r) \qquad (6.3)$$

where a, h are as earlier defined. This conditional distribution is such that the implied marginal $p(\boldsymbol{\theta}_{t+1}|D_t)$ has precisely the same mixture form as $p(\boldsymbol{\theta}_t|D_t)$, so that the local structure is respected. This theoretical discussion therefore indicates and opens up a direction for development that, if implemented, can be expected to generate more accurate and efficient methods of smoothing posterior samples. To exploit this mixture theory will require, among other things, computationally and statistically efficient methods of identifying the parameters R and $\{q_r, \boldsymbol{b}_r, \boldsymbol{B}_r : r = 1, \ldots, R\}$ of the mixture in equation (6.1) based on an existing Monte Carlo sample (and weights) from that distribution. Some form of hierarchical clustering of sample points (and weights), such as utilised in West (1993a,b), will be needed, though a key emphasis lies on computational efficiency so new clustering methods will be needed. Such developments, while challenging, will directly contribute in this context to usefully extend and improve the existing algorithms for sequential filtering on both parameters and states in higher dimensional dynamic models.

Acknowledgements

The authors are grateful for comments and fruitful discussions with Omar Aguilar, Simon Godsill, Neil Gordon, Pepe Quintana and an anonymous referee. This work was performed under partial support of NSF grant DMS-9704432 (USA), and while the first author was a PhD student at Duke University. The authors would also like to acknowledge the support of CDC Investment Management Corporation.

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation, *Journal of Business and Economic Statistics*.

Berzuini, C., Best, N. G., Gilks, W. R. and Larizza, C. (1997). Dynamic conditional independence models and Markov chain Monte Carlo methods, *Journal of the American Statistical Association* **92**: 1403–1412.

Doucet, A. (1998). On sequential simulation-based methods for Bayesian filtering, *Technical report CUED/F-INFENG/TR 310*, Department of Engineering, Cambridge University.

Geweke, J. F. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory, *The Review of Financial Studies* **9**: 557–587.

Givens, G. and Raftery, A. E. (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships, *Journal of the American Statistical Association* **91**: 132–141.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to non-linear/non-Gaussian Bayesian state estimation, *IEE Proceedings-F* **140**: 107–113.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting (with discussion), *Journal of the Royal Statistical Society, Series B* **38**: 205–247.

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with arch models, *Review of Economic Studies* **65**: 361–393.

Kitagawa, G. (1998). Self-organising state space model, *Journal of the American Statistical Association* **93**: 1203–1215.

Liu, J. S. and Chen, R. (1995). Sequential Monte Carlo methods for dynamic systems, *Journal of the American Statistical Association* **90**: 567–576.

Pitt, M. and Shephard, N. (1999a). Analysis of time varying covariances: A factor stochastic volatility approach, *in* J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 5*, University Press, Oxford, pp. 547–570.

Pitt, M. and Shephard, N. (1999b). Filtering via simulation: Auxiliary particle filters, *Journal of the American Statistical Association* **94**: 590–599.

Pole, A. (1988). Transfer response models: A numerical approach, *in* J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (eds), *Bayesian Statistics 3*, University Press, Oxford, pp. 733–746.

Pole, A. and West, M. (1990). Efficient Bayesian learning in non-linear dynamic models, *Journal of Forecasting* **9**: 119–136.

Pole, A., West, M. and Harrison, P. J. (1988). Non-normal and non-linear dynamic Bayesian modelling, *in* J. C. Spall (ed.), *Bayesian Analysis of Time Series and Dynamic Models*, Marcel Dekker, New York, pp. 167–198.

Smith, A. F. M. and West, M. (1983). Monitoring renal transplants: An application of the multi-process Kalman filter, *Biometrics* **39**: 867–878.

West, M. (1986). Bayesian model monitoring, *Journal of the Royal Statistical Society (Ser. B)* **48**: 70–78.

West, M. (1993a). Approximating posterior distributions by mixtures, *Journal of Royal Statistical Society* **55**: 409–422.

West, M. (1993b). Mixture models, Monte Carlo, Bayesian updating and dynamic models, *in* J. H. Newton (ed.), *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface*, Interface Foundation of North America, Fairfax Station, Virginia, pp. 325–333.

West, M. and Harrison, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models, *Journal of the American Statistical Association* **81**: 741–750.

West, M. and Harrison, P. J. (1989). Subjective intervention in formal models, *Journal of Forecasting* **8**: 33–53.

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd edn, Springer-Verlag, New York.