Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models

Paolo Giordani*
School of Economics
University of New South Wales

Robert Kohn†
School of Economics
School of Banking and Finance
University of New South Wales

May 2005

Abstract

Time series subject to parameter shifts of random magnitude and timing are commonly modeled with a change-point approach using Chib's (1998) algorithm to draw the break dates. We outline some advantages of an alternative approach in which breaks come through mixture distributions in state innovations, and for which the sampler of Gerlach, Carter and Kohn (2000) allows reliable and efficient inference. We show how the same sampler can be used to (i) model shifts in variance that occur independently of shifts in other parameters (ii) draw the break dates in $O(n)$ rather than $O(n^3)$ operations in the change-point model of Koop and Potter (2004b), the most flexible and general to date. Finally, we introduce to the time series literature the concept of adaptive Metropolis-Hastings sampling for discrete latent variable models. We develop an easily implemented adaptive algorithm that improves on Gerlach *et al.* (2000) and promises to significantly reduce computing time in a variety of problems including mixture innovation, change-point, regime-switching, and outlier detection. The efficiency gains on two models for U.S. inflation and real interest rates are 257% and 341%.

Keywords: Structural breaks; Parameter instability;
Change-point; State-space; Mixtures: Discrete latent variables; Adaptive Metropolis-Hastings.

JEL Classification Codes: C11, C15, C22.

*School of Economics, University of New South Wales, Sydney NSW 2052, Australia, e-mail: p.giordani@unsw.edu.au

†School of Economics, School of Banking and Finance, University of New South Wales, Australia, e-mail: r.kohn@unsw.edu.au

1 Introduction

This paper is concerned with the problem of modeling and inference for processes subjects to random shifts in parameters at unknown dates. The literature on the topic can be traced back at least to Quandt (1958) and Chernoff and Zack (1964), but has expanded rapidly in recent years because of faster computers and the availability of powerful new statistical tools for the solution of the considerable numerical problems involved. Additional incentives have been provided by Clements and Hendry (1999), who point to failures to detect (or to properly adjust to) intercept shifts as a main source of forecast failure, and by Stock and Watson (1996), who document the widespread nature of parameter instability in linear models of U.S. macroeconomic and financial time series.

Traditionally, a majority of papers in this area take a frequentist approach.[1] However, in recent years the number of Bayesian alternatives has increased rapidly, as many authors came to recognize the theoretical and practical advantages of a Bayesian approach for these types of problems. From a theoretical perspective, it seems desirable for parameter and forecast distributions to reflect uncertainty on the number and timing of break-points, rather than being conditioned on modal values. From a practical stance, because the likelihood quickly becomes intractable as the number of breaks increase[2], frequentists are forced to resort to simulation methods similar to those employed by Bayesians, therefore losing the advantage of simpler implementation enjoyed in other models.

The Bayesian approach coupled with modern simulation methods makes it tractable to estimate models with multiple structural breaks. Latent variables determine the location and nature of the breaks. Conditional on these latent variables, it is possible to generate the model parameters, while conditional on the parameters it is possible to generate the latent variables. Such generation is repeated until an adequate sample is obtained from the posterior distribution of interest. The most popular Bayesian approach in the econometric literature on structural breaks follows Chib (1998), who models the breaking process as a Markov chain with transition probabilities constrained so that regimes come in a non-reversible sequence. (We refer to this as a 'change-point' approach.) Pesaran, Pettenuzzo, and Timmermann (2004) extend the results in Chib (1998) by adding a hierarchical level for all parameters. Koop and Potter (2004a, 2004b) make a significant advance by allowing for an unknown number of breaks.

An alternative approach builds on the state-space representation, modeling

[1] The frequentist literature on testing for structural breaks is reviewed by Hansen (2001) and Elliott and Müller(2003).

[2] See the discussion in Elliott and Müller (2003).

the breaking process through mixture distributions for the state innovations. (We refer to this as a 'mixture innovation' approach.) Time-varying-parameter models with normal innovations in state variables are well-known and easily estimated special cases (see Cogley and Sargent (2001) and Primiceri (*forthcoming*) for recent Bayesian economic applications), but the general case of several component distributions presents formidable computational problems. Several approximations have been proposed, including Harrison and Stevens (1976), Sims (1993) and Engle and Smith (1999). Markov chain Monte Carlo techniques opened the way for the exact solutions developed by McCulloch and Tsay (1993), Carter and Kohn (1994), and Shephard (1994). However, these algorithm are at best extremely inefficient in the case of relatively infrequent breaks. The reason is that they draw the discrete latent variables conditional on the states, and in models with structural breaks the correlation between these two blocks is very high or even perfect. For models that can be written in conditionally Gaussian state-space form, the sampling algorithm of Gerlach, Carter and Kohn (2000) avoids this problem and permits fast and efficient estimation of mixture innovation models.

Our paper makes four contribution to the Bayesian literature on structural breaks. The first is to remark that mixture innovations in state variables are often an intuitive and flexible way of modelling breaks, and that even complex models can be estimated efficiently with the method of Gerlach, Carter and Kohn (2000). Mixture innovation models have a number of desirable features, including random number of in-sample breaks, possibility of jointly modeling small (and more frequent) and large (and less frequent) breaks, easy inclusion of innovation and additive outliers, ease of prior elicitation, and the relative convenience of allowing different parameters to change at different times.

The second contribution is to show how to perform efficient inference for models with breaks in conditional variance, and how to allow breaks in conditional mean parameters and in conditional variance to occur independently.

Of course in some instances a change-point approach may still be preferred. Koop and Potter (2004b) is then an interesting option because it allows for a random number of breaks. However, it requires $O(n^2)$ operations to draw the break dates with a geometric prior distribution for regime durations, and $O(n^3)$ operations with a Poisson prior (their preferred choice). The third contribution of this paper is to show that, for conditionally Gaussian models, the algorithm of Gerlach *et al.* (2000) can be adapted to draw the break-points in the model of Koop and Potter (2004b) in just $O(n)$ operations with either prior.

The fourth contribution is to introduce to the time series literature the concept of adaptive Metropolis-Hastings sampling for discrete latent variable mod-

els. We present a simple algorithm which further improves sampling efficiency (over Gerlach *et al.*, 2000) in a class of discrete latent variable models that includes change-point, mixture innovation, regime switching, and outlier detection. The adaptive algorithm uses past draws of the discrete latent variables to design a proposal distribution for a Metropolis-Hastings step, greatly reducing computing time spent (and mostly wasted) on observations for which the presence or absence of a break is rather clear-cut. In two applications to U.S. inflation and real interest rate, this adaptive algorithm reduces computing times (compared to Algorithm 1 in Gerlach *et al.* (2000), for a given Monte Carlo standard error) by over 75% and 60%.

The paper is organized as follows. Section 2 presents three Bayesian approaches to structural change, and discusses some advantages of mixture innovation models. In particular, it argues that the method of Gerlach *et al.* (2000) solves the computational problems that have so far constrained their use and development. Section 4 suggests a method to model breaks in variances in conditionally Gaussian form and to combine shifts in conditional mean and variance parameters. Section 4 shows how the sampler of Gerlach *et al.* (2000) can be adapted to draw the break dates in models with non-Markov breaking processes, using the case of a Poisson prior for regime duration as an example. Section 5 outlines a new adaptive Metropolis-Hastings sampler which promises to further increase sampling efficiency in a variety of discrete latent variable problems. Section 6 evaluates the efficiency gains from the new adaptive sampler on models of U.S. inflation and real interest rate. Section 7 concludes.

2 Bayesian approaches to structural change

This section reviews three approaches to modeling structural change: change-point models, time-varying-parameter models, and mixture innovation models. For our purposes, it is useful to cast all these models in a common framework, writing them in state-space form, with system matrices depending on vectors of discrete latent variables. The contributions of our article are for conditionally Gaussian models. Conditionally Gaussian models, introduced independently by Carter and Kohn (1994) and Shephard (1994), can be written in linear and Gaussian state-space form conditional on a vector of latent variables:

$$y_t = g_t + h_t' x_t + \gamma_t u_t \tag{1}$$

$$x_t = f_t + F_t x_{t-1} + \Gamma_t v_t, \tag{2}$$

where $u_t \sim nid(0,1)$, $v_t \sim nid(0,I)$[3] and all system matrices are known (for $t = 1, ..., n$) conditional on parameters, observations and on discrete latent variables K, where, throughout the paper, for any variable z_t, we define $z^{t_1,t_2} = (z_{t_1}, ..., z_{t_2})$ and $z = (z_1, ..., z_n)$. The latent variables are typically assumed to be independent or Markov. Most time series models take a conditionally Gaussian state-space form, and need not be Gaussian after integrating out the latent variables. In fact, besides models for structural change, the most common nonlinear models and several models for fat-tailed errors are conditionally Gaussian (see Giordani, Kohn and van Dijk, 2005).

2.1 Change-point models with a known number of breaks

A conditionally linear change-point model with a known number of breaks is defined as

$$\begin{aligned} y_t &= \beta_m^{'} z_t + \sigma_m u_t \\ \beta_m &= \beta_1, \ \sigma_m = \sigma_1, \ \text{for } \tau_1 > t \geq 1 \\ &... \\ \beta_m &= \beta_{M-1}, \ \sigma_m = \sigma_{M-1}, \ \text{for } \tau_{M-1} > t \geq \tau_{M-2} \\ \beta_m &= \beta_M, \sigma_m = \sigma_M, \ \text{for } n \geq t \geq \tau_{M-1}, \end{aligned} \tag{3}$$

where $u_t \sim nid(0,1)$, M-1 is the number of breaks and the unknown parameters are $(\beta_1, ..., \beta_M, \sigma_1, ..., \sigma_M, \tau_1, ..., \tau_{M-1})$. Chib (1998) obtains large efficiency gains over previous samplers by framing the change-point problem as a switching-regime problem with transition probabilities constrained so that regimes come in a non-reversible sequence. This model can be written is the state-space form (1)-(2) by setting $K_t = m$, $g_t = \beta^{'}(K_t) z_t$, $\gamma_t = \sigma(K_t)$, and modelling K_t as a discrete Markov process with restrictions on the transition probability matrix. Chib then adapts the sampler of Carter and Kohn (1994) and draws K in $O(nM)$ operations. He then shows how to compute the marginal likelihood so that the number of breaks can be treated as an unknown quantity in a model selection or model averaging context. Chib's method does not require the measurement equation to be linear or Gaussian, but it does require the probability of y_t to be readily available given parameters $\theta(K_t)$, and is therefore impractical for models with time-varying unobserved states such as measurement errors or seasonal components.

[3] It is possible to allow $E(u_t v_t) \neq 0$.

2.2 Change-point models with an unknown number of breaks

Pesaran, Pettenuzzo and Timmermann (2004) generalize the model of Chib (1998), but retain the assumption of a known number of breaks. Even though they integrate over the number of breaks M-1 by computing the marginal likelihood for $M = 0, 1, ..., M_{max}$, Koop and Potter (2004a) point out that there are several drawbacks to this strategy, besides the large computational costs. One is the difficulty in nesting the case of rare, large breaks, and of frequent, smaller breaks. A second problem is that the need to impose exactly M-1 in-sample breaks typically implies a prior in which breaks near the end of the sample are inordinately more likely than at earlier dates. A further problem is that computing the marginal likelihoods accurately in complex models is very difficult.

To avoid these shortcomings, Koop and Potter (2004b) develop a change-point model with an unknown number of breaks. Letting the number of breaks be random is not only valuable per se, but also delivers a prior in which all observations have the same probability of being a break-point. Their key idea is to model an unknown number of breaks indirectly, by assuming that there are M regimes but that up to M-1 may occur out of sample. They set $M = n$ and are therefore able to nest a time-varying-parameter model and a model with few breaks (given a sufficiently sparse prior on regime duration).

Another innovation in Koop and Potter (2004b) is that regime duration has a Poisson prior in place of the standard geometric prior. A geometric prior for regime duration corresponds to the assumption of an independent breaking process, i.e. $p(K_t|K_{s\neq t}) = p(K_t)$, where $K_t = 1$ if there is a break at time t and $K_t = 0$ otherwise. The geometric distribution implies that, for any $j > 1$, durations of j are more likely than durations of $j + 1$, a feature of the prior that is relaxed in Koop and Potter (2004b).[4]

Koop and Potter (2004b) draw the break dates conditional on parameters by adapting the algorithm of Chib (1998). Chib's algorithm, originally designed for a geometric prior on duration, requires $O(nM)$ operations to draw all break points. Koop and Potter set $M = n$, resulting in $O(n^2)$ operations. Transition

[4] Koop and Potter (2004a) also show that prior regime duration remains approximately geometric if K_t is Markov. The idea of departing from a geometric distribution is not uncontroversial, though. Gelman *et al.* (1995, page 52) find that the geometric distribution's "'memoryless' property makes it a natural model for survival or lifetime data". Moreover, we note that the use of a Poisson prior may be questionable when breaks are thought to be (or turn out to be) relatively rare. For example, suppose that breaks occur every twenty years on average. On monthly data that corresponds to $\lambda = 240$, which implies a standard deviation of $\sqrt{240} \simeq 15.5$, or a little over a year. The Poisson prior therefore restricts infrequent breaks to recur with greater regularity than is probably prudent to assume.

probabilities are constant for a geometric distribution, whereas they depend on the distance from the last break for a Poisson. The choice of a Poisson prior therefore requires averaging over durations in computing transition probabilities, leading to $O(n^3)$ operations. This is unfortunate, since in practice it limits the use of what is otherwise a flexible and powerful model. Section 4 shows how break points can be drawn in $O(n)$ operations by adapting the sampling algorithm of Gerlach *et al.* (2000) rather than that of Chib (1998).

2.3 Time varying parameter models

Time-varying-parameter (TVP) models are written in the state-space form (1)-(2) with constant system matrices. Parameter variation comes through normal innovations in the transition equations. These models have a long tradition (West and Harrison (1997) is a classic Bayesian reference). Recent examples in economics include Cogley and Sargent (2001) and Primiceri (*forthcoming*). Inference by Gibbs sampling is relatively straightforward, as a number of algorithms are available to draw the states x conditional on parameters (including Carter and Kohn (1994), Fruhwirth-Schnatter (1994), de Jong and Shepard (1995) and Durbin and Koopman, 2002). The limitation of TVP models is that they are designed for smooth and frequent parameter variation and cannot effectively cope with infrequent interventions. Both types of breaks can be captured by mixture innovation models.

2.4 Mixture innovation models

Mixture innovation models are very general because they allow all system matrices in (1) and (2) to depend on K_t. Parameter shifts are typically modeled by letting $\Gamma_t\Gamma_t^{'}$ (the covariance matrix of innovations to the states) depend on K_t. To illustrate, consider the simple example

$$\begin{aligned} y_t &= \mu_t + \sigma_u u_t \\ \mu_t &= \mu_{t-1} + K_t \sigma_v v_t, \end{aligned} \tag{4}$$

where K_t is an independent sequence of Bernoulli variables, with

$$\begin{aligned} K_t &= 1 \text{ with probability } \pi \\ K_t &= 0 \text{ with probability } 1-\pi, \end{aligned}$$

and u_t and v_t are both $nid(0,1)$. This model can be written in terms of (1) and (2) by setting $g_t = 0$, $h_t = 1$, $\gamma_t = \sigma_u$, $f_t = 0$, $F_t = 1$, $\Gamma_t = \sigma_v K_t$. At each new observation, there is a probability π of a break; if there is a break,

the change in the mean of the process is normally distributed with mean zero and standard deviation σ_v. The number of in-sample breaks is random, with prior distribution defined by π (possibly a random variable itself). Prior regime duration is geometric, with mean $(1-\pi)/\pi$ and standard deviation $\sqrt{(1-\pi)}/\pi$.[5]

This approach to structural change has a number of advantages over the change-point approach:

1. It is an intuitively simple way of allowing for random number and size of breaks, which facilitates prior elicitation.

2. Computational costs are linear in the number of values that K_t can take. This gives the modeler great flexibility. For example, we may wish to capture the idea that the mean of a process is unchanged in some periods, while in others it incurs small but relatively frequent and large but infrequent breaks. This can be achieved by setting

$$\sigma_v(K_t = K^i) = \sigma_{v_i} \ , \ prob(K_t = K^i) = \pi_i \text{ for } i = 1, 2, 3,$$

with $\sigma_{v_1} = 0$, and a prior for which $\sigma_{v_3} \gg \sigma_{v_2}$ and $\pi_2 \gg \pi_3$. This is a general and intuitive way of nesting TVP models and models with infrequent breaks.[6]

3. It is just as easy to assume that parameters in different regimes (i) are drawn from a common distribution, as in Chib (1998) and Pesaran *et al.* (2004) (ii) depend only on parameters in the previous regime, as in Koop and Potter (2004b) and in most TVP and mixture innovation applications, or (iii) are a mixture of (i) and (ii). Referring to equation (4), the first case can be modelled as

$$\mu_t = mK_t + \mu_{t-1}(1 - K_t) + K_t\sigma_v v_t, \tag{5}$$

where K_t is either 0 or 1. To nest both cases, write

$$\mu_t = mK_{1t} + \mu_{t-1}(1 - K_{1t}) + (K_{1t}\sigma_{1v} + K_{2t}\sigma_{2v})v_t, \tag{6}$$

where $K_t = (K_{1t}, K_{2t})$, $K_t = (0, 0)$ for no break, and $K_t = (1, 0)$ and $(0, 1)$ for a break with $\mu_t|K_t$ independent and a random walk respectively.

[5] Notice that a geometric distribution with mean μ has standard deviation close to μ, whereas a Poisson distribution with mean μ has standard deviation $\sqrt{\mu}$. When μ is relatively large, the two priors have very different implications for the regularity of occurrences of breaks.

[6] Koop and Potter (2004b) also nest a TVP model and a model with few breaks, given a sufficiently sparse prior on durations. However, mixture innovations are more general in that they allow to combine (rather than choose from) these two options.

4. It is a convenient framework to allow different parameters to change at different times. Change-point models invariably assume that all parameters change at the same time (unless some are forced to be constant). Without this restriction, a change-point model would have M^p regimes to keep track of, where M-1 is the maximum number of breaks in each parameter and p is the number of parameters that are allowed to change, implying at least $O(nM^p)$ operations to draw the break dates. In contrast, if we let $K_{it} \in \{0, 1\}$ denote the absence or presence of a break in parameter i at time t, then $K_t = (K_{1t}, ..., K_{pt})$ can take 2^p values, so K can be drawn in $O(n2^p)$ operations regardless of the number of breaks (see Section 2.4.1).

5. It is a convenient framework to allow shifts in variance and in conditional mean parameters to occur independently (see Section 3).

6. It is a convenient framework to model breaks and outliers jointly. This can be important, as innovation and (especially) additive outliers may both hide actual breaks and spuriously indicate non-existing breaks, particularly in real-time forecasting.

7. While sampling schemes for multiple change-point models draw the break dates conditional on all parameters, in a mixture innovation approach most parameters can be treated as states and integrated out, which increases sampling efficiency. For example, referring to the model given by (3), we would write

$$\begin{aligned} y_t &= \beta_t^{'} z_t + \sigma_t u_t \\ \beta_t &= \beta_{t-1} + K_t \sigma_v v_t, \end{aligned}$$

   treat β_t as the state vector and integrate it out when drawing K_t (see Section 2.4.1).

Finally, it is worth noting that mixture innovations can be combined with regime-switching and change-points (see Giordani *et al.* (2005) for an example), and that transition and break probabilities can be related to exogenous variables (see McCulloch and Tsay, 1993).

2.4.1 Sampling K in mixture innovation models

Mixture innovation models have been the subject of considerable study[7], but until the mid-90s exact solutions were available only for the simplest cases and at

[7] Harrison and Stevens (1976) call them multiprocess models. See also Kitagawa (1987).

great computational expense. Several approximations were proposed, including Harrison and Stevens (1976), Sims (1993) and Engle and Smith (1999). Markov chain Monte Carlo techniques opened the way for the exact solutions developed by McCulloch and Tsay (1993), Carter and Kohn (1994), and Shephard (1994). However, the algorithms proposed by Carter and Kohn (1994) and Shephard (1994) draw the auxiliary variables K conditional on the states x. In the case of structural breaks or additive outliers, K and x are highly correlated, making these samplers very inefficient. Often, as in the model given in (4), the correlation is perfect, and the samplers break down completely (see Gerlach *et al.*, 2000). McCulloch and Tsay (1993) reduce the severity of this problem by drawing K_t conditional on the error v_t rather than on the state x_t, but their sampler remains rather inefficient (Gerlach *et al.*, 2000). Gerlach *et al.* (2000) develop an algorithm for conditionally Gaussian processes that draws K without conditioning on x, and therefore retains a high degree of efficiency regardless of the correlation between K and x.

The first contribution of Gerlach *et al.* (2000) is to draw K from

$$p(K_t|y, K_{s\neq t}, \theta) \propto p(y^{t,n}|y^{1,t-1}, K, \theta)p(K_t|K_{s\neq t}, \theta), \tag{7}$$

where the states have been integrated out (rather than conditioned on) and θ is the vector of parameters. For a given proposed value of K_t, $p(K_t|K_{s\neq t}, \theta)$ is evaluated from the transition probabilities, and $p(y^{t,n}|y^{1,t-1}, K, \theta)$ can be computed with the Kalman filter in conditionally Gaussian models. Evaluating $p(y^{t,n}|y^{1,t-1}, K, \theta)$ through the Kalman filter is straightforward, but requires $O(n)$ operations, implying $O(n^2)$ operations to draw K. Gerlach *et al.* (2000)'s second contribution is to provide an algorithm to evaluate $p(y^{t,n}|y^{1,t-1}, K, \theta)$ in one step and thus to draw K in $O(n)$ operations. Since K_t takes a finite number of values, it can be drawn by computing $p(K_t|y, K_{s\neq t}, \theta)$ for all possible values of K_t and then normalizing. The computing time is then linear in both n and in the number of elements in the mixture (the number of values that K_t can take).

3 Breaks in conditional variance

A mixture innovation approach can also be used to model variance shifts, as first shown by McCallum and Tsay (1993). Their algorithm draws K conditional on the states and is therefore inefficient. Drawing K without conditioning on states as in Gerlach *et al.* (2000) is preferable. However, because squared residuals are not normal, it would seem that their algorithm is not suitable for variance shifts. However, this problem is overcome as follows. We begin by considering

shifts in variance in an otherwise white noise process, and then show how to combine shifts in variances with shifts in other parameters.

Modeling shifts in variance. Let y_t be a zero-mean, uncorrelated random variable such that

$$y_t = \sigma_t \epsilon_t, \tag{8}$$

where ϵ_t is standard normal. Then

$$log(y_t^2) = log(\sigma_t^2) + u_t, \tag{9}$$

where u_t is $log(\chi_1^2)$. Following the stochastic volatility literature, we work with $log(\sigma_t^2)$, ensuring that σ_t^2 is always positive. It is then natural to model permanent shifts in σ_t^2 as

$$log(\sigma_t^2) = log(\sigma_{t-1}^2) + \sigma_v(K_{2,t})v_t, \tag{10}$$

where v_t is a standard normal random variable. If K_{2t} takes only one value the model reduces to smoothly changing variances of the type used by Primiceri (*forthcoming*). In Section 6 we set $\sigma_v(K_{2,t} = 0) = 0$ and $\sigma_v(K_{2,t} = 1) = \sigma_v^* > 0$ with a prior favouring infrequent shifts. More elements can be added to the mixture, for example to allow for both smooth and sudden shifts in variance, analogously to the case of shifts in conditional mean parameters.

To draw K_{2t}, we follow Carter and Kohn (1993 and 1997), Shephard (1994) and Kim *et al.* (1998), who observe that the distribution of a $log(\chi_1^2)$ can be very accurately approximated by a mixture of normals with few components. (Carter and Kohn (1997) use five, Kim *et al.* (1998) seven.[8]) We can then write the model in conditionally Gaussian state-space form

$$\begin{aligned} z_t &= g(K_{1,t}) + h_t + G(K_{1,t})\eta_t \\ h_t &= h_{t-1} + \sigma_v(K_{2,t})v_t, \end{aligned} \tag{11}$$

where $z_t = log(y_t^2)$, $h_t = log(\sigma_t^2)$, and η_t and v_t are standard normal. The mean and standard deviation of each component of the mixture approximating the distribution of u_t determine $g(K_{1,t})$ and $G(K_{1,t})$ respectively, so these matrices contain no unknown parameters. The prior probabilities π_i of each component are also known. Thus

$$p(u_t) \simeq \sum_i^I \pi_i N(g_i, G_i^2),$$

[8] Both papers report all parameters of the mixture. We use the parameters from Carter and Kohn (1997).

where $N(g_i, G_i^2)$ is the normal density with mean g_i and variance G_i^2.

Shifts in variance can then be modeled through $\sigma_v(K_{2,t})$. If K_{2t} can take only one value the model reduces to smoothly changing variances of the type used by Primiceri (*forthcoming*). In Section 6 we set $\sigma_v(K_{2,t} = 0) = 0$ and $\sigma_v(K_{2,t} = 1) = \sigma_v^* > 0$ with a prior of infrequent shifts. More elements can be added to the mixture, for example to allow for both smooth and sudden shifts in variance, analogously to the case of shifts in conditional mean parameters.

Shifts in conditional first and second moments. We use the following sampling scheme to deal with shifts in both the conditional mean and conditional variance parameters:

1. Initialize the sampler with a time series of conditional standard deviations $\sigma^{1,n}$ (for example $\sigma_1 = \sigma_2 = ... = \sigma_n$).

2. Given $\sigma^{1,n}$, the model is conditionally Gaussian. Draw interventions K_m as in Gerlach *et al.* (2000). Conditional on K_m, draw the states x with any of the algorithms mentioned in Section 2.3. Conditional on y, K_m and x, compute residuals r_t. The definition of residuals is such that $std(r_t) = \sigma_t$.

3. Model interventions in $log(r_t^2)$ as explained above in this section. Referring to equation (11), draw $K_v = (K_1^{1,n}, K_2^{1,n})$. Given K_v, draw $log\left(\sigma_t^2\right)$ for $t = 1, .., n$ with the same algorithms used to draw x in step (2).

4. Go to (2).

In going from (2) to (3), the most convenient assumption is that that shifts in conditional variance (second block) are independent of shifts in conditional mean parameters (first block). It is possible to have a prior relating probabilities of interventions in the first and second block, but some care is required as an overly tight prior may induce a nearly reducible chain. Innovation and additive outliers are drawn in the first block and accounted for in computing the residuals used in the second block (see Section 6 for an example), therefore reducing the risk of isolated outliers being intrerpreted as variance shifts.

4 Non-Markov breaking processes

This sections shows how to adapt the algorithm of Gerlach *et al.* (2000) when K_t is not Markov, and how this allows drawing the break dates in $O(n)$ operations in the change-point model of Koop and Potter (2004b) with a Poisson prior on regime durations.

Although Gerlach *et al.* (2000) work with Markov interventions, their key results do not require this assumption. In particular, the algorithm to draw $p(y^{t,n}|K)$ in one step does not rely on K_t being Markov, as all operations are conditioned on K, and it is therefore suitable for any breaking process as long as it is possible to evaluate $p(K_t|K_{s\neq t},\theta)$ (or $p(K|\theta) \propto p(K_t|K_{s\neq t},,\theta)$). The independence and Markov assumptions are simply the most convenient to perform this evaluation.

4.1 Poisson prior for regime duration

Following Koop and Potter (2004b), assume that regime duration follow a Poisson distribution. The duration of regime m is defined as $d_m = \tau_{m+1} - \tau_m$, where τ_{m+1} and τ_m are adjacent break dates. Then they assume that $d_m - 1$ is Poisson distributed $Po(\lambda)$ with parameter λ.[9] We now show how to draw K in mixture innovation models with a Poisson prior for durations, and then how the same technique can be used to draw break dates in the model of Koop and Potter (2004b). In both cases, only $O(n)$ operations are required for conditionally Gaussian models.

4.1.1 Drawing K

Let $K_t = 1$ ($K_t = 0$) stand for the presence (absence) of a break at time t. We want to evaluate $p(K_t|K_{s\neq t})$—where the dependence on parameters θ has been dropped for convenience—when durations are Poisson distributed with known parameter λ. Assuming that there is at least one break both before and after t, the probability of a break at t is given by

$$p(K_t = 1|K_{s\neq t}) = p(d_{1t})p(d_{2t}),$$

where d_{1t} and d_{2t} are regime durations, defined as the number of periods between two adjacent breaks

$$\begin{aligned} d_{1t} &= min(t - \tau_1;\ t > \tau_1, K_{\tau_1} = 1) \\ d_{2t} &= min(\tau_2 - t;\ t < \tau_2, K_{\tau_2} = 1), \end{aligned}$$

and the probability of no break at t is given by

$$p(K_t = 0|K_{s\neq t}) = p(d_{3t}),$$

[9] Koop and Potter (2004b) use a hierarchial prior to allow the mean duration λ to depend on m. For our purposes, we can assume a single λ with no loss of generality.

where d_{3t} is the number of periods between the last break before period t and the first break after period t

$$d_{3t} = min(\tau_2 - \tau_1;\ \tau_2 > t > \tau_1,\ K_{\tau_1} = K_{\tau_2} = 1).$$

Finally, $p(d_{it})$, where $i = 1, 2, 3$, is evaluated as

$$p(d_{it}) \propto \frac{exp(-\lambda)\lambda^{d_{it}-1}}{(d_{it}-1)!}.$$

The only slight complication arise when there is no in-sample break either before or after t. A simple but unsatisfactory solution is to assume $K_0 = K_{n+1} = 1$. To illustrate a better solution, keep the assumption $K_0 = 1$ (simply for convenience of exposition). Initialize the algorithm with an arbitrary value of n^*, such as $n^* = n + 1$, where

$$n^* = min(\tau;\ \tau > n,\ K_\tau = 1).$$

Given n^*, sequentially update $K_1, ..., K_n$ as described above. n^* is then easily updated from the prior conditional on $n - \tau_l$, where such as stochastic volatility models is the last in-sample break date.

4.1.2 Application to Koop and Potter (2004b)

Referring to equation (7), having shown how to evaluate $p(K_t|y, K_{s\neq t}, \theta)$ with a Poisson prior on duration, we can use Gerlach *et al.* (2000) to compute $p(y^{t,n}|K, \theta)$ in one step if the model is conditionally Gaussian.[10] To illustrate, consider a simplified version of the benchmark model in Koop and Potter (2004b):

$$\begin{aligned} y_t &= \beta_m^{'} x_t + \sigma_m u_t \\ \beta_m &= \beta_{m-1} + \eta_t^{\beta} \\ ln(\sigma_m) &= ln(\sigma_{m-1}) + \eta_t^{\sigma} \\ var(\epsilon_t) &= 1,\ var(\eta_t^{\beta}) = \Omega_\beta,\ var(\eta_t^{\sigma}) = \omega_\sigma, \end{aligned}$$

where $d_m - 1 \sim Po(\lambda)$ and other priors can be omitted for our purposes. Koop and Potter draw all parameters conditional on the regimes and then the regimes conditional on the parameter sub-set $\theta = (\beta_1, ..., \beta_m, \sigma_1, ..., \sigma_m, \lambda)$. If we frame the problem as one of drawing break dates rather than regimes, then, conditional on θ and on the break dates, the model can be written as (1)-(2) with all system matrices known, by setting

$$g_t = \beta_m^{'} x_t,\ h_t = 0,\ \gamma_t = \sigma_m.$$

[10] Non-Gaussian change-point models include Chib (1998) and Koop and Potter (2004a).

This completes the description of how to draw K in $O(n)$ operations using the sampler of Gerlach *et al.* (2000).

5 Adaptive Metropolis-Hastings sampling for discrete latent variable models

5.1 Discussion

This section describes a new adaptive sampler that speeds up the drawing of K. Since this is, to the best of our knowledge, the first application of adaptive sampling for discrete latent variables in a time series context, the main ideas are discussed in some detail. Because the discussion is in terms of drawing K in a conditionally Gaussian model, the methods developed in this section are applicable to mixture innovation models as well as regime switching and multiple change-point models with unobserved states and to the change-point model with unknown number of breaks of Koop and Potter (2004b). Moreover, they are ideally suited to detect outliers in both time series and cross sectional data.

The key idea of adaptive sampling is to use previous draws to form efficient proposal distributions for the application of the Metropolis-Hastings method. Precursors to full adaptive sampling limited this learning process to a subset of the burn-in period. This guarantees that the chain is Markov, and hence that the standard convergence properties of Metropolis-Hastings are unaffected by the initial learning period. A fully adaptive sampler, however, does not limit the learning to the burn-in period. Because the proposal distribution is no longer constant, but rather depends on the history of the draws, the resulting chain is not Markov, so the usual proof of ergodicity of the Metropolis-Hastings algorithm no longer applies. Establishing the properties of an adaptive MH sampler is a rather complex problem. Recent progress has been made by Gilks, Roberts and Sahu (1998), Haario, Saksamn and Tamminen (2001), Atachade' and Rosenthal (2003), and Nott and Kohn (*forthcoming*).

In particular, Nott and Kohn (*forthcoming*) establish conditions for the validity of an adaptive sampling scheme for discrete distributions. However, their argument is also valid for a compact space and so can be applied by suitably truncating the priors on the unknown parameters and states without affecting practical performance. Adaptation needs to be implemented with care to ensure ergodicity. Loosely speaking, the main requirements are that we use the entire history of draws rather than a moving window and that we constrain the proposal distribution so that no event in the sample space of K has zero

probability.

The traditional use of partially (limited to the burn-in) adaptive schemes is to gain some knowledge of a non-standard distribution. Adaptive algorithms can of course be used to increase sampling efficiency in this case. However, the intuition of Nott and Kohn (*forthcoming*), which we extend to our time-series framework, is that an adaptive Metropolis-Hastings algorithm can increase sampling efficiency even when direct sampling from the distribution of interest is possible. The key idea is to use information coming from the history of the draws to cheaply draw from an approximation rather than expensively draw from the exact distribution. The approximation is then used as the proposal distribution in a Metropolis-Hastings step.

The general problem of detecting regimes and interventions in time series models is ideally suited to incorporate these ideas, because (i) the distribution of $(K_t|Y, K_{s\neq t})$ is expensive to compute, particularly when the number of interventions is random (ii) in most periods, most of the probability mass is concentrated in a small sub-set of the sample space of K_t. For example, during the running of the chain, for many observations it soon becomes clear whether a break in that period is likely or unlikely, or if the observation belongs to, say, regime one or two. Sizable computational gains are therefore possible by incorporating this information into an adaptive algorithm, rather than drawing K_t from the exact but computationally expensive distribution.

5.2 Sampling algorithm.

We now present our adaptive Metropolis-Hastings (MH) sampling algorithm to draw K efficiently. The algorithm is first stated and then discussed.

1. For $m = 1, ..., d$, generate $K^{(m)}$ (the m-th draw of K) using the algorithm of Gerlach *et al.* (2000). After drawing each $K^{(m)}$, update all the model parameters.

2. For $m = d + 1, ..., M$ (where M is the total number of iterations), carry out (a)-(c) for $t = 1, ..., n$, and then (d).

   (a) Using past draws $K^{(1)}, ..., K^{(m-2)}$, compute the percentage of $K_t = K^i$ for $i = 1, ..., J$, where J is the number of values that K_t can take. Call each percentage $p_{i,t}$.

   (b) Define $\alpha_{it} = max(min(p_{it}, 1-\delta), \delta)$, for $i = 1, ..., J$, where $\delta > 0$ (e.g. 0.01 or 0.02). Then normalize α_{it} so that $\sum_{i=1}^{J} \alpha_{it} = 1$. Draw $P \in \{1, ..., J\}$ from a multinomial density with parameters $\{\alpha_{1t}, ..., \alpha_{Jt}\}$. Given P, form K^P.

(c) Let C be the current indicator of K_t, that is, $K^C = K_t^{(m-1)}$. We accept $K_t^{(m)} = K^P$ with probability $min(1, p)$, where

$$p = \frac{p(K_t = K^P|y, K_{s\neq t})}{p(K_t = K^C|y, K_{s\neq t})} \frac{\alpha_{Ct}}{\alpha_{Pt}},$$

and $p(K_t = K^P|y, K_{s\neq t})$ can be computed efficiently using the method of Gerlach *et al.* (2000), and (crucially) no computation is required whenever $K^P = K^C$.

(d) Update all model parameters.

We note that adaptation is carried out only on the latent variables K_t and not on the other parameters and variables in the model.

The learning mechanism outlined in step 2.a has two characteristics that help ensure ergodicity: (i) an expanding (rather than fixed) window of draws (ii) a proposal distribution built with information up to m-2 only ($K^{(m-1)}$ is available but not used). The percentages mentioned in step 2.a may be updated at each iteration or only infrequently (we update them every 50 iterations). In step 2.b, a strictly positive constant δ ensures that we can draw with positive probability even those interventions that have not been drawn before (thus ensuring that the sampling scheme does not become reducible). Crucially, for a large proportion of the observations the multinomial density $(\alpha_{1t}, ..., \alpha_{Jt})$ typically assigns most of the probability mass to one or a few outcomes. For example, for most observations the probability of an outlier is either close to zero or to one. The probability of a break at t is rarely close to one (exception for very large breaks), but is often very small. The proposal distribution built in 2.b is used in 2.c in a Metropolis-Hastings step.

5.3 Assessing efficiency gains

The efficiency gains of our adaptive MH sampler are likely to be larger when interventions are relatively rare (as for breaks) and/or relatively clear-cut (as for outliers and, in some cases, regime-switching), since then the large majority of proposals is of the type $K^P = K^C$, and hence no calculation is required in step 2.*c*. Computational gains are also likely to increase when J is large. To say more, we need to specify a model, a data-set and a prior, as the efficiency of our adaptive MH sampler can only be evaluated numerically.

Since the number of objects of interest (including K) is high, we summarize sample draws through the distributions of selected parameters and of the states at $t = 1$. Denoting by θ_i^m the m-th draw of parameter i, we compare Algorithm 1 in Gerlach *et al.* (2000) with our adaptive MH sampler by (i) confirming that

they lead to the same posterior inference, summarized by $\overline{\theta_i} = (\sum_{m=1}^{M} \theta_i^m)/M$ and $\widehat{\sigma_i}^2 = [\sum_{m=1}^{M} (\theta_i^m - \overline{\theta_i})^2]/M$, and (ii) comparing their relative efficiency . (ii) is done as follows: An estimate of the variance of $\overline{\theta_i}$ is

$$var(\overline{\theta_i}) = \frac{\widehat{\sigma_i}^2}{M}[1 + 2\sum_{i=1}^{R}(1 - \frac{j}{M})\widehat{\rho}(j)],$$

where $\widehat{\rho}(j) = corr(\widehat{\theta^m, \theta^{m-j}})$ and the truncation point R is such that $\widehat{\rho}(j) \simeq 0$ for $j > R$. Assuming that M is sufficiently large, $\widehat{\sigma_i}^2$ is the same for both samplers. The term in square parenthesis is the inefficiency factor (IF). An independent sampler has unit inefficiency factor. We can couple the inefficiency factor with the running time of the two sampling schemes to obtain a measure of the relative efficiency of the adaptive scheme. The number of iterations needed for the adaptive scheme to have the same $var(\overline{\theta_i})$ as the non-adaptive scheme is given by the ratio of the inefficiency factors. We can then multiply this number by the ratio of computing time (CT) to obtain a measure of the relative efficiency of the adaptive Metropolis-Hastings sampler:

$$\text{Relative efficiency} = \frac{\text{CT non-adaptive}}{\text{CT adaptive}} \frac{\text{IF non-adaptive}}{\text{IF adaptive}}.$$

6 Efficiency gains in models of U.S. inflation and real interest rate

This section computes relative efficiencies in two applications to U.S. real interest rate and inflation. The models are meant to illustrate the mixture innovation approach to structural breaks and to evaluate the efficiency of the adaptive sampler rather than to provide a thorough empirical analysis of the two series.

In both cases 10000 iterations are used, with a burn-in of 100.[11] The first 100 iterations are always performed with Algorithm 1 in Gerlach *et al.* (2000). After 100 iterations we either continue to use the same algorithm or switch to the adaptive Metropolis-Hastings sampler, with $\delta = 0.01$.

6.1 Mean shifts in U.S. real interest rates

U.S. real interest rates have attracted considerable attention in the multiple change-point literature because they seem to display several sharp shifts in mean, at least in a univariate context (Bai and Perron, 1998). Our data are

[11] The burn-in may seem low, but in our experience convergence to the stationary distribution occurs very quickly in this type of model using Algorithm 1 in Gerlach *et al.* (2000).

quarterly from 1952Q1 to 2004Q4[12] and consider an AR(1) model with fixed autoregressive coefficient and fixed residual variance, time variation modeled directly in the mean of the process rather than in the intercept, and additive outliers

$$\begin{aligned} y_t &= y_t^* + \sigma_e K_{v,t} v_t \\ y_t^* &= \mu_t + \rho(y_{t-1} - \mu_{t-1}) + \sigma_e e_t \\ \mu_t &= \mu_{t-1} + \sigma_e K_{\mu,t} u_t \\ p(K_t) &= p(K_t | K_{s \neq t}), \end{aligned} \tag{12}$$

where $K_t = (K_{v,t}, K_{\mu,t})$. Additive outliers are modelled by letting $K_{v,t} \in \{0, g_v\}$ and breaks in mean by letting $K_{\mu,t} \in \{0, g_{\mu_1}, g_{\mu_2}\}$, where we now allow $(g_v, g_{\mu_1}, g_{\mu_2})$ to be random. In total, K_t can take the four values (for given $g_v, g_{\mu_1}, g_{\mu_2}$), shown in Table 1.

The priors for $(g_v^2, g_{\mu_1}^2, g_{\mu_2}^2)$ are inverse gamma

$$g_i^2 \sim IG(S_i, n_i),$$

where $i \in \{v, \mu_1, \mu_2\}$, $n_i = 5$ and $\sqrt{S_v}/n_v = 2.5$, $\sqrt{S_{\mu_1}}/n_{\mu_1} = 1$, $\sqrt{S_{\mu 2}}/n_{\mu_2} = 3$. These priors are not very tight, but they do not try to be diffuse either. As shown in McCulloch (2000), being uninformative or nearly uninformative on the size of breaks ends up placing most prior probability on extremely large breaks, with the result that no break is ever found. Conversely, diffuse priors on the frequency of breaks place considerable probability on very frequent breaks, and therefore carry a large risk of risk of over-fitting.

Letting π_i be the prior probability that $K_t = K^i$, the prior distribution for $(\pi_1, \pi_2, \pi_3, \pi_4)$ is Dirichelet

$$p(\pi_1, \pi_2, \pi_3, \pi_4) \sim D(n_0 \pi_1^*, ..., n_0 \pi_4^*),$$

where n_0 can be interpreted as the number of prior observations (the tightness of the prior), and $(\pi_1^*, ..., \pi_4^*)$ as their relative frequencies (the prior mean).[13] To emphasize the strength of the evidence of mean shifts, we impose a tight prior of extremely infrequent breaks: $n_0 = 1000$, and the prior probabilities

[12] The nominal interest rate is the yield on three months government bonds, secondary market. Inflation is CPI inflation for urban consumers, all items, seasonally adjusted. Both series are from the database FRED II (http://research.stlouisfed.org/fred2), aggregated from monthly data.

[13] The posterior distribution of $g_v^2, g_{\mu_1}^2, g_{\mu,}^2$ given σ_ϵ is inverse gamma. See Giordani *et al.* (2005) for details.

of $K_\mu = g_{\mu_1}$ and $K_\mu = g_{\mu 2}$ are both just 0.1%, implying an average regime duration of 125 years. In spite of such tight priors, posterior distributions show rather clear evidence of four mean shifts, corresponding closely with common historical interpretations of the data: low real interest rate during the inflation of the 70s, then a sudden increase with the Volcker deflation, and after a few years a return to lower values. Once mean shifts are taken into account, the real interest rate displays little persistence: Table 1 shows that the posterior mean of the autoregressive coefficient ρ is only 0.35.

Table 3 shows that the relative efficiencies of the MH sampler for μ_1, ρ, σ_e are 3.13, 3.64, and 3.47 respectively. That is, on average, the adaptive sampling scheme reaches a given Monte Carlo standard error in about 29% of the time required by Algorithm 1 in Gerlach *et al.* (2000). The adaptive scheme is a bit less efficient per iteration, but the efficiency loss is compensated for by the increase in speed.

6.2 Changing mean, dynamics and volatility in U.S. inflation

Several studies (see, for example, Nelson and Piger, 2002) have suggested the presence of structural breaks in the mean of U.S. inflation.[14] Moreover, there is some statistical evidence and some economic theory to support the non-constancy of inflation persistence over long time periods (Akerlof *et al.* (2000), Christiano and Fitzgerald, 2003). However, other researchers have found little variation in either the intercept or autoregressive coefficients of AR models for U.S. inflation, but strong evidence of changes in residual variance (for example, Primiceri (*forthcoming*), and Koop and Potter, 2004b).

We model quarterly U.S. CPI[15] inflation for the period 1951Q1-2004Q4 as an AR(1) process with random breaks in intercept, autoregressive parameter, and residual variance as:

[14] This statement implictly assumes a linear reference model. Sargent (1999) is an example of a non-linear model in which shifts in the mean and dynamics of inflation arise endogenously, a reminder that the presence or absence of structural breaks can only be evaluated in relation to a model.

[15] Consumer Price Index for all urban consumers (all items), from the database Fred II, series ID: CPIAUCSL, seasonally adjusted. CPI aggregated from monthly data (averages) and inflation defined as $400(\frac{CPI_t}{CPI_{t-1}} - 1)$.

$$
\begin{aligned}
y_t &= c_t + b_t y_{t-1} + \sigma_t K_{e,t} e_t \\
c_t &= c_{t-1} + \sigma_t K_{c,t} u_t^c \\
b_t &= b_{t-1} + \sigma_t K_{b,t} u_t^b \\
log(\sigma_t^2) &= log(\sigma_{t-1}^2) + K_{v,t} v_t \\
p(K_t) &\equiv p(K_{m,t}, K_{v,t}) = p(K_{m,t}) p(K_{v,t}) = p(K_t | K_{s \neq t}),
\end{aligned} \tag{13}
$$

where $K_{m,t} = (K_{e,t}, K_{c,t}, K_{b,t})$. The latent variable $K_{e,t}$ can take values $(1, 2.5)$, where 1 is a standard observation and 2.5 an innovation outlier; $K_{c,t}$ can take the values $(0, 0.2, 1)$ and $K_{b,t}$ can take values $(0, 0.5)$. For ease of interpretation, we assume that a break and an outlier cannot occur simultaneously, but we do allow breaks in c_t and b_t to occur both separately and jointly. This last feature seems new to the literature. As discussed in Section 2.4, extending change-point models to relax the assumption that all parameters break at the same time is straightforward in principle, but problematic in practice, whereas in our case the increased computational burden is small. In total, the vector $K_{m,t}$ can take seven values (refer to Table 2). $K_{v,t}$ can take the values (0, 1.39); $\sigma_t = \sigma_{t-1}$ for $K_{v,t} = 0$, while $K_{v,t} = 1.39$ and $v_t = 1$ (-1) imply $\sigma_t / \sigma_{t-1} \simeq 2$ $(\sigma_t / \sigma_{t-1} \simeq 0.5)$.

The prior probabilities of interventions are fixed and reflect the assumption that breaks in any parameter are rare (the combined probability of a break in c_t and/or b_t is 1%, so the prior mean interval between breaks is 25 years; the probability of a break in variance is also 1%). Priors on all the other parameters are conditionally conjugate but diffuse, and centered on OLS values. Sampling states and parameters conditional on K is then straightforward (see Giordani *et al.* (2005) for details).

The results are summarized in Figure **??**. The rise in inflation in the late sixties and seventies is captured by an increase in both coefficients, whereas the falling inflation of the eighties leaves the constant nearly unaffected but corresponds to a large and sudden decrease in persistence. Inflation persistence is a positive function of the inflation level, as observed by Akerlof *et al.* (2000). The conditional standard deviation shows marked changes during the sample, which are also strongly correlated with the inflation level.[16]

Table 3 shows that the relative efficiencies of the MH sampler for c_1, b_1, σ_1 are 1.94, 2.82, and 2.94 respectively.

[16] This finding is hardly surprising. There is conclusive evidence that inflation is less forecastable at higher levels (see Giordani and Söderlind (2003) for a review of the literature).

7 Conclusion

Koop and Potter (2004a and 2004b) convincingly illustrate the advantages of allowing for a random number of breaks in change-point models. Our article argues that a mixture innovation approach is often a more natural and general way of achieving the same goal, and that fast and reliable estimation of conditionally Gaussian mixture innovation models is possible using the algorithm of Gerlach *et al.* (2000). The sampler of Chib (1998) is efficient when the number of break-points is known, but can otherwise involve dramatically higher computational costs.

Our article also introduces to the time series literature the concept of adaptive Metropolis-Hastings sampling for discrete latent variable problems. A simple adaptive scheme produces large efficiency gains (arising from reduced computing times) in two empirical applications to models with outliers and shifts in both conditional mean and conditional variance parameters. When considering the generality and simplicity of the sampler and the small additional programming costs required, these efficiency gains suggest that adaptive Metropolis-Hastings for discrete latent variable models should prove a fruitful concept in time series.

Acknowledgments. *The work of Robert Kohn was partially supported by an Australian Research Council Grant on mixture models.*

References

[1] Akerlof, G., Dickens, W.T., and Perry, G.L., 2000, Near rational wage and price setting and the optimal rates of inflation and unemployment, *Brookings Papers on Economic Activity*, 1, 1-60.

[2] Albert, J.H., and Chib, S. 1993, Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts, J*ournal of Business and Economic Statistics*, 11, 1-15.

[3] Atachade', Y., and Rosenthal, J., 2003. On adaptive Markov chain Monte Carlo algorithms, Technical Report, Department of Mathematics and Statistics, University of Montreal.

[4] Bai, J., and Perron, P., 1998, Estimating and testing linear models with multiple structural changes, *Econometrica*, 66, 47-78.

[5] Barnett, G., Kohn, R., and Sheather, S. 1996, Bayesian estimation of an autoregressive model using Markov chain Monte Carlo, *Journal of Econometrics*, 74, 237-254.

[6] Carter, C. K., and Kohn, R., 1993, On the applications of Markov chain Monte Carlo methods to linear state space models, in Proc. Bus. Econ. Statist. pp. 131-136.

[7] Carter, C.K., and Kohn, R., 1994, On Gibbs sampling for state-space models, *Biometrika*, 83, 589-601.

[8] Carter, C.K., and Kohn, R., 1997, Semiparametric Bayesian inference for time series with mixed spectra, *Journal of the Royal Statistical Society*, Series B, 255-68.

[9] Chernoff, H., and Zacks, S. 1964, Estimating the current mean of a normal distribution which is subject to changes in time, *Annals of Mathematical Statistics*, 35, 999-1028.

[10] Chib, S. 1998, Estimation and comparison of multiple change-point models, *Journal of Econometrics*, 86, 221-241.

[11] Christiano, L.J, and Fitzgerald, T.J., 2003, Inflation and monetary policy in the twentieth century, *Chicago Fed Economic Perspectives*, 1, 1-24.

[12] Clements, M.P., and Hendry, D.F., 1999, *Forecasting non-stationary economic time series*, Cambridge, Mass., MIT Press.

[13] de Jong, P. and Shephard, N., 1995, The simulation smoother for time series models, *Biometrika*, 82, 339-350.

[14] Cogley, T. and Sargent, T., 2001, Evolving post-World War II inflation dynamics, *NBER Macroeconomic Annual*.

[15] Durbin, J., and Koopman, S., 2002, A simple and efficient simulation smoother for state space time series analysis, *Biometrika*, 89, 603-616.

[16] Elliott, G., and Müller, U.K., 2003, Optimally testing general breaking processes in linear time series models, Working Paper, Princeton University.

[17] Fruhwirth-Schnatter, S., 1994, Data augmentation and dynamic linear models, *Journal of Time Series Analysis*, 15, 183-202.

[18] Gilks, W.R., Richardson, S. and Sahu, S.K. 1998, Adaptive Markov chain Monte Carlo through regeneration, J*ournal of the American Statistical Association*, 93, 1045-1054.

[19] Geweke, J. 2003, *Contemporary Bayesian econometrics and statistics*, Manuscript, University of Iowa.

[20] Giordani, P., and P. Söderlind, 2003, Inflation forecast uncertainty, European Economic Review, 47, 1037-59.

[21] Giordani, P., Kohn, R., and van Dijk, D., 2005, A unified approach to nonlinearity, structural change and outliers, Manuscript, University of New South Wales.

[22] Haario, H., Saksman, E. and Tamminen, J. 2001, An adaptive Metropolis algorithm, *Bernoulli*, 7, 223-242.

[23] Hamilton, J.D. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, 57, 357-84.

[24] Hansen, B., 2001, The new econometrics of structural change: Dating breaks in U.S. labor productivity, *Journal of Economic Pespectives*, 15, 4, 117-128.

[25] Harrison, J., and Stevens, C.F. 1976. Bayesian forecasting (with discussion), *Journal of the Royal Statistical Association*, B 38, 205-47.

[26] Kim, C.J., and Nelson, C.R. 1998, Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime-switching, *Review of Economics and Economic Statistics*, 80, 88-101.

[27] Kim, S., Shephard, N., and Chib, S., 1998, Stochastic volatility: Likelihood inference and comparison with ARCH models, *Review of Economic Studies*, 65, 361-393.

[28] Kitagawa, G., 1987. Non-Gaussian state space modeling of non-stationary time series, *Journal of the American Statistical Association*, 82, 1032-41.

[29] Koop, G., and Potter, S.M., 2004a, Prior elicitation in multiple change-point models, Manuscript, University of Leicester.

[30] Koop, G., and Potter, S.M., 2004b, Forecasting and estimating multiple change-point models with an unknown number of change-points, Manuscript, University of Leicester.

[31] Koopman, S.J., and Rubin, J. 2000, Fast filtering and smoothing for multivariate state space models, *Journal of Time Series Analysis*, 21, 281-296.

[32] Levin, A.T., and Nelson, C.R., 2003, Is inflation persistence intrinsic in industrial economies? Federal Reserve Bank of St. Louis Working Paper 2002-023E.

[33] McCulloch, R.E., and Tsay, R.S., 1993, Bayesian inference and prediction for mean and variance shifts in autoregressive time series, *Journal of the American Statistical Association*, 88, 968-978.

[34] McCulloch, J.H. 2000, State-space time series modeling of structural breaks, Working Paper number 00-11, Ohio State University.

[35] Nott, D., and Kohn, R., *forthcoming*, Adaptive sampling for Bayesian variable selection, *Biometrika*.

[36] Pesaran, H., Pettenuzzo, D., and Timmermann, A. 2004, Forecasting time series subject to multiple structural breaks, Manuscript, University of California, San Diego.

[37] Primiceri, G., *forthcoming*, Time varying structural vector autoregressions and monetary policy, *Review of Economic Studies*.

[38] Quandt, R., 1958, The estimation of the parameters of a linear regression system obeying two separate regimes, *Journal of the American Statistical Association*, 53, 873-880.

[39] Sargent, T.J., 1999, *The conquest of American inflation*, Princeton, N.J., Princeton University Press

[40] Shephard, N.E, 1994, Partial non-Gaussian state-space models, *Biometrika*, 81, 115-131.

[41] Stock, J., and Watson, M., 1996, Evidence on structural instability in macroeconomic time series relations, *Journal of Business and Economic Statistics*, 14, 11-30.

[42] West, M., and Harrison, J., 1997, *Bayesian forecasting and dynamic models*, Springer-Verlag, New York.

	K^1	K^2	K^3	K^4
$K_{v,t}$	0	g_v	0	0
$K_{\mu,t}$	0	0	$g_{\mu 1}$	$g_{\mu 2}$
prior prob	0.978	0.020	0.001	0.001
post. prob	0.975	0.019	0.002	0.004
	$E(\cdot\|y)$	$std(\cdot\|y)$		
g_v	2.48	0.82		
g_{μ_1}	1.19	0.53		
$g_{\mu 2}$	3.57	1.52		
ρ	0.35	0.08		

Table 1: **Values of K(t) and probabilities for model of the U.S. real interest rate.** Values of K(t) in columns. First row: value of Kv (additive outlier). Second row: value of Km (break in mean). Third row: prior probability of each value of K(t). Fourth row: posterior probability of each value of K(t). Further rows: mean and std of model parameters.

	K^1	K^2	K^3	K^4	K^5	K^6	K^7
$K_{e,t}$	1	2.5	1	1	1	1	1
$K_{c,t}$	0	0	0.2	0.2	1	0	1
$K_{b,t}$	0	0	0	0.5	0	0.5	0.5
prob	0.97	0.02	0.002	0.002	0.002	0.002	0.002

Table 2: **Values of Km(t) and probabilities for AR(1) model of U.S. inflation.** Values of K(t) in columns. First row: value of Ke (innovation outlier). Second row: value of Kc (break in constant). Third row: value of Kb (break in autoregressive parameter). Kv(t) can take values 0 and 1.39 with prior probabilities 0.99 and 0.01, independent of Ke, Kc, Kb.

Inflation	IF	IF ratio	RE		**Real rate**	IF	IF ratio	RE
c_1	3.32	1.76	1.94		μ_1	1.26	1.14	3.13
b_1	2.14	1.21	2.82		ρ	1.68	0.98	3.64
σ_1	2.89	1.16	2.94		σ	1.58	1.03	3.47

Table 3: **Inefficiency factor of the adaptive MH algorithm, the ratio of the inefficiency factors of the adaptive MH algorithm over the algorithm of Gerlach et al. (2000), and the relative efficiency for selected parameters in two models of U.S. inflation and real interest rate.**

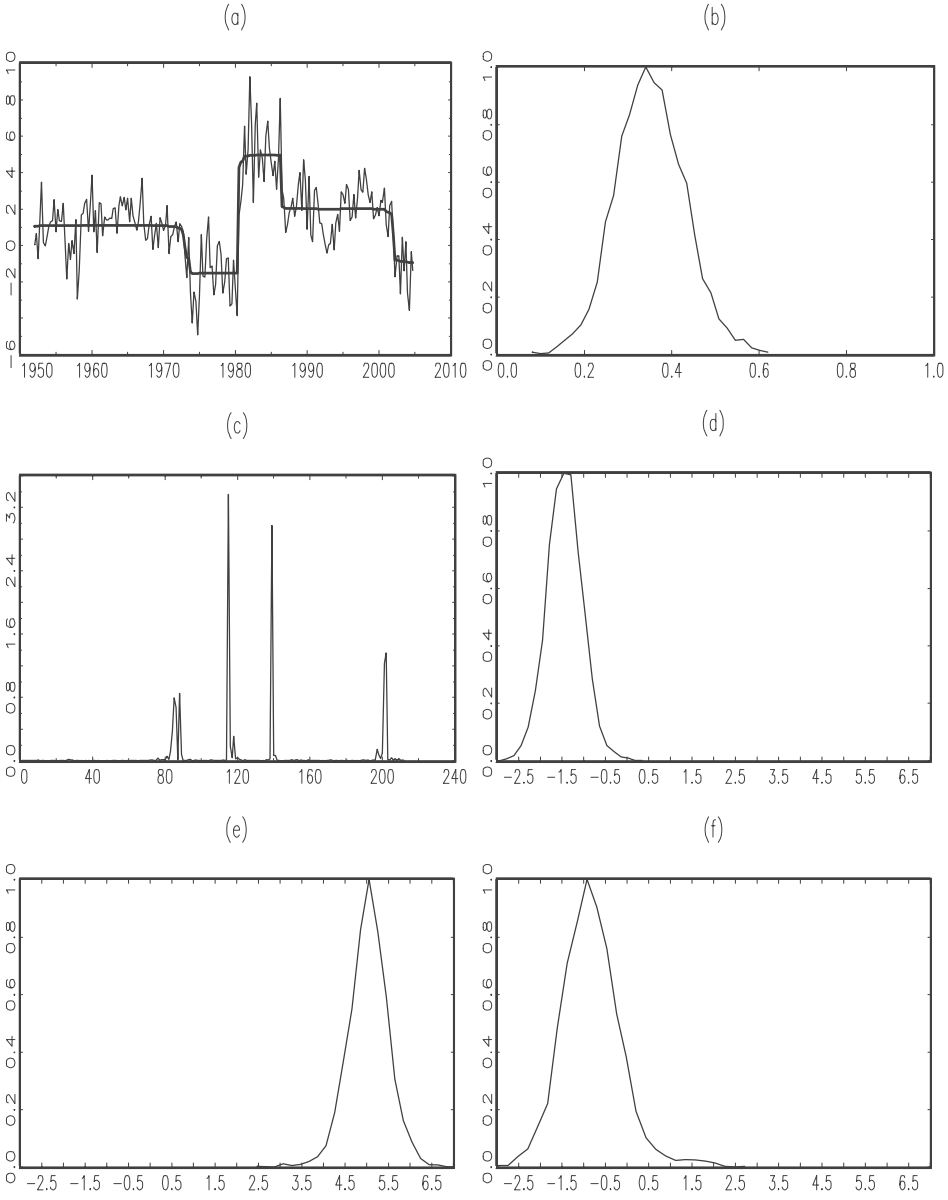

Figure 1: AR(1) model with mean shifts for the U.S. real interest rate. (a) real interest rate and posterior mean of μ_t (b) posterior density of ρ (c) posterior mean of $K_{m,t}$ (d) posterior distribution of μ_{1976Q1} (e) posterior distribution of $\mu_{19822Q1}$ (f) posterior distribution of μ_{2004Q4}.

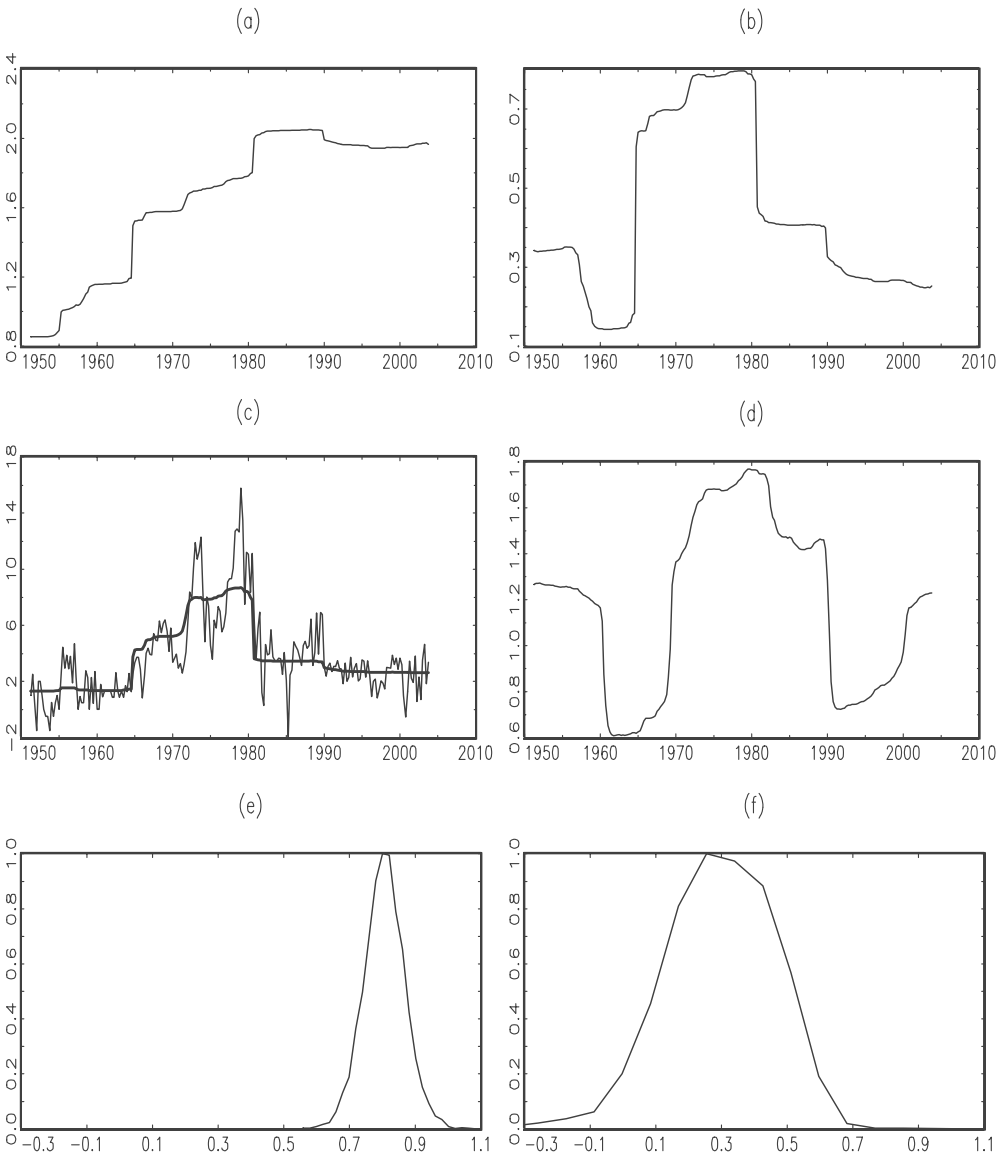

Figure 2: AR(1) model of U.S. inflation: (a) posterior mean of c_t (b) posterior mean of b_t (c) inflation and posterior median of $c_t/(1-b_t)$ (d) posterior mean of σ_t (e) posterior distribution of b_{1979Q1} (f) posterior distribution of b_{2004Q4}.