Geweke, John

Article

Sequentially adaptive Bayesian learning for a nonlinear model of the secular and cyclical behavior of US real GDP

Econometrics

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Geweke, John (2016) : Sequentially adaptive Bayesian learning for a nonlinear model of the secular and cyclical behavior of US real GDP, Econometrics, ISSN 2225-1146, MDPI, Basel, Vol. 4, Iss. 1, pp. 1-23, https://doi.org/10.3390/econometrics4010010

This Version is available at:
https://hdl.handle.net/10419/171861

econometrics

MDPI

Article

Sequentially Adaptive Bayesian Learning for a Nonlinear Model of the Secular and Cyclical Behavior of US Real GDP

John Geweke

Economics Discipline Group, School of Business, University of Technology Sydney, 14 - 28 Ultimo Road, Ultimo, NSW 2007, Australia; John.Geweke@uts.edu.au; Tel.: +61-2-9514-9797

Academic Editors: Herman K. van Dijk, Francesco Ravazzolo, Nalan Basturk and Roberto Casarin
Received: 1 October 2015; Accepted: 2 February 2016; Published: 2 March 2016

Abstract: There is a one-to-one mapping between the conventional time series parameters of a third-order autoregression and the more interpretable parameters of secular half-life, cyclical half-life and cycle period. The latter parameterization is better suited to interpretation of results using both Bayesian and maximum likelihood methods and to expression of a substantive prior distribution using Bayesian methods. The paper demonstrates how to approach both problems using the sequentially adaptive Bayesian learning algorithm and sequentially adaptive Bayesian learning algorithm (SABL) software, which eliminates virtually of the substantial technical overhead required in conventional approaches and produces results quickly and reliably. The work utilizes methodological innovations in SABL including optimization of irregular and multimodal functions and production of the conventional maximum likelihood asymptotic variance matrix as a by-product.

Keywords: business cycles; posterior simulation; sequential Monte Carlo

JEL classification: C11

1. Introduction

This paper takes up a simple but nontrivial example of a situation that arises often in econometrics. An econometric model takes the form $p(y \mid x, \beta)$ where y is a vector of outcomes, x is a vector or matrix of covariates or conditioning observables, β is an unknown parameter vector, and the functional form of p is the model specification. Due to the motivating economic theory, or by virtue of the interpretation of the model, β is a function of a more fundamental parameter vector θ, so that $\beta = \beta(\theta)$. The function need not be one-to-one. For purposes of motivation and interpretation θ is superior to β; indeed, this typically can best be accomplished in terms of θ. Examples can be found in most chapters of many econometrics textbooks, from the undergraduate to postgraduate levels. A generic case is the one in which θ expresses the underlying taste and/or technology state of a market or economy: neoclassical models of production and consumption and structural simultaneous equation macroeconometric models are two broad subcategories. The same is true in nonstructural settings like the many specific variants of linear state space models and factor models. And it is also true in the effort to provide economic interpretation of simple descriptive models like the one used here.

For a subjective Bayesian, this point has added force because a prior distribution that is substantive (by implication proper) must be expressed in terms of θ, not β. Bayesian econometrics now has a set of tools to address these nonlinear models, the most widely applied perhaps being Markov chain Monte Carlo (MCMC). This paper adds to this collection of tools the sequentially

adaptive Bayesian learning algorithm (SABL). Given that there already exist multiple approaches, the question of why an additional approach naturally arises. There are several good answers to this question. The objective of this paper is to substantiate the answers, by both presenting the method and by illustrating its application to nonlinear models.

1. SABL provides a generic and robust approach to the nonlinear model problem as just stated. Compared with existing methods like MCMC it requires very little tuning and experimentation with the exact form of the algorithm (for example, specification of the variance matrix for the Metropolis random walk, burn-in and convergence decisions). In many cases, like the illustration here, no tuning or experimentation at all is required.
2. With a simple change of a single parameter, the SABL algorithm can be used for optimization as well as Bayesian inference. Here, that feature is used to compute maximum likelihood estimates and their associated asymptotic distribution. More generally, SABL can be used to attack a wide variety of optimization problems in economics.
3. SABL enables Bayesians to use almost any prior distribution and to modify conventional distribution by imposing constraints (illustrated here) or mixing with discrete distributions (not illustrated). It also provides numerical standard errors and accurate log marginal likelihood (marginal data density) approximations as by-products, neither of which is a focus of this paper.
4. In many cases SABL is faster and more accurate than any competing algorithm when executed in a conventional multicore CPU environment. That is the case for the illustration here.
5. The SABL algorithm is pleasingly parallel, and except for a computationally insignificant portion of the algorithm it is embarrassingly parallel. It is therefore well suited to execution on graphics processing units, and this facility is included in SABL.

This list is not intended to be exhaustive, but rather as motivation for the reader already steeped in Bayesian computational techniques to continue on.

The paper proceeds by introducing the nonlinear model used as a simple but realistic example, Section 2. Section 3 then provides an overview of SABL used for Bayesian inference, drawing on more extensive documentation for SABL that is readily available [1]. This is followed by the illustrative example for Bayesian inference, Section 4. Section 5 takes up optimization in SABL, with a specific focus on maximum likelihood, and the illustration follows in Section 6. There is a short concluding section.

The paper is written to convey an understanding of SABL and its application to the problem set forth in the next section. For full development of the immediate theory, refer to [2–4]. For the underlying probability theory the single best self-contained reference is [5].

2. Interpretation of the Third-Order Autoregression

The relationship of second-order stochastic difference equations to business cycle dynamics has been recognized at least since [6]. In a model like

$$y_t = \gamma_0 + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \varepsilon_t,$$

where ε_t is the innovation (shock) for the stationary time series $\{y_t\}$, the associated characteristic polynomial is $\gamma(z) = 1 - \gamma_1 z - \gamma z^2$. Let the roots of the polynomial be r_1 and r_2. Stationarity is equivalent to $|r_1| > 1$, $|r_2| > 1$. If, in addition, r_1 and r_2 are a complex conjugate pair then the time series exhibits characteristic cycles in which the amplitude is $\alpha = |r_1|^{-1} = |r_2|^{-1}$ and the period is $2\pi / \tan^{-1}(Im(r_1)/Re(r_1))$ where $\tan^{-1}$ denotes the principle branch in $(0, \pi)$. This is the technical essence of [6] and now a standard part of graduate education.

Geweke [7] applied a similar interpretation to the third-order stochastic difference equation (autogregression)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} N\left(0, \sigma^2\right), \tag{1}$$

taking up the case with one real root r_1 and a pair of complex conjugate roots r_2 and r_3 in the associated characteristic polynomial. That paper used the reparameterization

$$\alpha_s = |r_1|^{-1},\ \alpha_c = |r_2|^{-1},\ p = 2\pi / \tan^{-1}\left(\mathrm{Im}\left(r_2\right) / \mathrm{Re}\left(r_2\right)\right), \tag{2}$$

where the subscript s denotes "secular" and c denotes "cyclical." It continued to impose $\alpha_c < 1$, permitted $\alpha_s > 1$, $\alpha_s - 1$ then being the explosive growth rate, as well as $\alpha_s < 1$, the rate of dampening in the transmission of ε_t. The paper used a conventional improper prior distribution for β and $\sigma^2 = \mathrm{var}\left(\varepsilon_t\right)$ and data from 19 OECD countries 1957–1983. It found posterior probabilities of explosive roots near 0, probabilities of complex pairs exceeding one-half for all countries, and probabilities that $|a_s| > |\alpha_c|$ over one-half for most countries.

The application here works directly with parameters that are much closer to the way that economists think about growth and business cycles than are the time series parameters $\beta = \left(\beta_0, \beta_1, \beta_2, \beta_3\right)'$ and σ. This enables the economist to construct a substantive prior distribution and to readily understand the posterior distribution.

The work begins by replacing the amplitudes α_s and α_c with their corresponding half-lives. A damping amplitude $\alpha \in (0,1)$ generates the sequence $1, \alpha, \alpha^2, \ldots$, equivalent to the impulse response function in the first-order autoregression $x_t = \alpha x_{t-1} + \eta_t$. The corresponding half-life is the value h for which

$$\frac{\int_0^h \alpha^u du}{\int_0^\infty \alpha^u du} = \frac{1}{2},$$

which implies $h = \log\left(1/2\right) / \log\left(\alpha\right)$. Thus for the secular and cyclical components we have the inverse relations

$$\alpha_s = (1/2)^{1/h_s},\ \alpha_c = (1/2)^{1/h_c} \tag{3}$$

mapping the secular half-life h_s to its amplitude α_s and the cyclical half-life h_c to its amplitude α_c. Writing the generating polynomial for the stochastic difference equation (1)

$$1 - \beta_1 z - \beta_2 z^2 - \beta_3 z^3 = \left(1 - r_1 z\right)\left(1 - r_2 z\right)\left(1 - r_3 z\right)$$

with $r_3 = \bar{r}_2$, we obtain from (2)

$$\beta_1 = \alpha_s + 2\alpha_c \cos\left(2\pi / p\right),\ \beta_2 = -\left(\alpha_s \alpha_c\right) \cos\left(2\pi / p\right) + \alpha_c^2,\ \beta_3 = \alpha_s \alpha_c^2. \tag{4}$$

The domains of α_s and α_c are $(0, \infty)$. The domain of p is $(2, \infty)$, the lower bound corresponding to using the principal branch of the function $\tan^{-1}$ in $(0, \pi)$ in (2). More fundamentally, there is an identification problem presented by aliasing: in any time series with one time unit between observations periodicities $p/\left(1 + pj\right)$ $(j = 0, 1, 2, \ldots)$ all exhibit as period p. A conventional way to resolve this problem is to require $p > 2$, which also seems satisfactory for business cycles with annual data.

Thus, given the half-lives h_s and h_c and the period p, (3) followed by (4) provides $\beta_1, \beta_2, \beta_3$. The mapping is continuous and has continuous derivatives of all orders. The other parameters of the model are σ, which has straightforward economic interpretation, and β_0, which does not but is of little interest.

It is natural to express independent prior distributions for h_s, h_c, p, and σ: most economists could readily state some plausible and some implausible values for each—unlike the case with β_1, β_2 and β_3. The half-lives and periods are all measured in time units and the supports of the respective prior distributions must each be (a subset of) the positive half-line. Several well-known prior distributions qualify and can easily be used in the approach taken here. We will use a log-normal prior distribution

for each, the distribution for $\log(p)$ being truncated below at $\log 2$. There are also several well-known prior distributions for σ; we again use a log normal prior. Finally, β_0 will be given a diffuse normal prior distribution.

Both the Bayesian approach (Sections 3 and 4) and the maximum likelihood approach (Sections 5 and 6) thus use the 5×1 parameter vector θ with

$$\beta_0 = \theta_1, h_s = \exp(\theta_2), h_c = \exp(\theta_3), p = \exp(\theta_4), \sigma = \exp(\theta_5).$$

The components $\theta_1, \ldots, \theta_5$ are mutually independent and normal in the prior distribution. Section 4 discusses the specific choice of the distribution. The prior distribution is also used as a computational device for maximum likelihood, but the result is unaffected by the choice of prior so long as its support includes the maximizing argument of the likelihood function. Noting that the successive transformations (2) to (3) to (4) are continuously differentiable of all orders, and considering the properties of the likelihood function for (1), it is clear that if the likelihood function has a unique internal global mode, then there is a neighborhood of the mode in which the log-likelihood function is twice continuously differentiable with bounded third derivative. This familiar condition is important to a deep verification of the properties of the SABL maximum likelihood estimates in Sections 5 and 6.

3. Bayesian Inference Using SABL

The SABL algorithm is a procedure for the controlled introduction of new information. It pertains to situations in which information can be represented as the probability distribution of a finite dimensional vector. SABL approximates this distribution by means of many (typically on the order of 10^4 to 10^6) alternative versions of the vector. These versions are called *particles*, reflecting some of SABL's connections to the particle filtering literature. In the SABL algorithm particles undergo a sequence of transformations as information is introduced. With minor exceptions accounting for a negligible fraction of computing time in typical research applications, these transformations amount to identical instructions that operate on each particle in isolation. SABL is therefore a pleasingly parallel algorithm. This property is responsible for dramatic decreases in computing time for many research applications with GPU execution of SABL.

At its highest level the SABL algorithm looks like this:

- Represent initial information
- While information not entirely incorporated
  - Determine information increment and incorporate by weighting particles
  - Remove the weights by resampling
  - Modify the particles to represent the information more efficiently
- End

In the sequential Monte Carlo literature each pass through the loop While ... End is known as a *cycle*, and we will use ℓ to index cycles. The three steps in each cycle are the *correction (C) phase*, the *selection (S) phase*, and the *mutation (M) phase*.

Let $\theta \in \Theta \subseteq \mathbb{R}^d$ denote the vector whose probability distribution represents information. Denote the particles by θ_{jn}, the double subscripts indicating the J groups of N particles each employed by SABL. Initially θ has probability density $p^0(\theta)$; extension beyond absolutely continuous distributions is easy, and this streamlines the notation. In SABL the particles initially are

$$\theta_{jn}^{(0)} \overset{iid}{\sim} p^{(0)}(\theta) \quad (j = 1, \ldots, J; n = 1, \ldots, N). \tag{5}$$

In Bayesian inference $p^{(0)}(\theta)$ is a proper prior density and in optimization it is the probability density. It must be practical to sample from the initial distribution (5) and to evaluate $p^{(0)}(\theta)$.

Denote the density incorporating all the information by $p^*(\theta)$. SABL requires that it be possible to evaluate a kernel $k(\theta)$ with the properties

$$k(\theta) \geq 0 \; \forall \; \theta \in \Theta, \quad \int_{\Theta} k(\theta) \, d\theta < \infty, \quad p^*(\theta) \propto k^*(\theta) = p^{(0)}(\theta) \, k(\theta). \tag{6}$$

In Bayesian inference the kernel $k(\theta)$ is the likelihood function,

$$k(\theta) = p(y_{1:T} \mid \theta), \tag{7}$$

where T denotes sample size and $y_{1:T} = \{y_1, \ldots, y_T\}$ denotes the data.

Cycle ℓ begins with the kernel $k^{(\ell-1)}$ and ends with the kernel $k^{(\ell)}$. In the first and last cycles,

$$k^{(0)} = 1 \quad \text{and} \quad k^{(L)}(\theta) = k(\theta),$$

respectively. Correspondingly define

$$k^{*(\ell)}(\theta) = p^{(0)}(\theta) \, k^{(\ell)}(\theta), \tag{8}$$

implying

$$k^{*(0)} = p^{(0)}(\theta) \quad \text{and} \quad k^{*(L)}(\theta) = k^*(\theta). \tag{9}$$

The particles change in each cycle, and reflecting this let $\theta_{jn}^{(\ell)}$ denote the particles at the end of cycle ℓ. The initial particles $\theta_{jn}^{(0)}$ have the common distribution (5) and are independent. In succeeding cycles the particles $\theta_{jn}^{(\ell)}$ continue to be identically distributed but they are not independent. The theory underlying SABL, discussed further in this section and developed in detail by [3,4] drawing on sequential Monte Carlo theory, assures that the final particles $\theta_{jn} = \theta_{jn}^{(L)} \xrightarrow{d} p^*(\theta)$. This convergence in distribution takes place in N, the number of particles per group. The result is actually stronger: the particles are ergodic in N, meaning that for any function g for which $E[g(\theta)] = \int_{\Theta} g(\theta) \, p^*(\theta) \, d\theta$ exists,

$$\lim_{N \to \infty} N^{-1} \sum_{n=1}^{N} g(\theta_{jn}) = E[g(\theta)] \tag{10}$$

with probability 1 in each group $j = 1, \ldots, J$.

A leading technical challenge in practical sequential Monte Carlo algorithms, which of course work with finite N, is to limit the dependence amongst particles, and in particular to keep dependence from increasing from one cycle to the next to the point that the final distribution of particles is an unreliable representation of any distribution at all. A further technical challenge is to provide a measure of the accuracy of the approximation implicit in the left side of (10) for finite N that is itself reliable. The SABL algorithm and toolbox do both in a way that makes minimal demands on users. The remainder of this section provides some details.

3.1. C phase

For each cycle ℓ define the weight function

$$w^{(\ell)}(\theta) = k^{(\ell)}(\theta) / k^{(\ell-1)}(\theta).$$

The theory underlying the SABL algorithm requires that there exist an upper bound $\overline{w}^{(\ell)}$, that is,

$$w^{(\ell)}(\theta) < \overline{w}^{(\ell)} < \infty \; \forall \; \theta \in \Theta.$$

The C phase determines $w^{(\ell)}(\theta)$ explicitly and thereby defines

$$k^{(\ell)}(\theta) = w^{(\ell)}(\theta) \cdot k^{(\ell-1)}(\theta) \tag{11}$$

and

$$p^{*(\ell)}(\theta) = k^{*(\ell)}(\theta)\, d\theta / \int_{\Theta} k^{*(\ell)}(\theta)\, d\theta.$$

Thus (8) and (11) imply $k^{*(\ell)}(\theta) = w^{(\ell)}(\theta) \cdot k^{*(\ell-1)}(\theta)$ as well. In SABL the weight functions $w^{(\ell)}(\theta)$ are designed so that there exists $L < \infty$ for which $k^{(L)}(\theta) = k(\theta)$, although the value of L is in general not known at the outset.

One approach in designing the weight function is to use the functional form $w^{(\ell)}(\theta) = k(\theta)^{\Delta_\ell}$ and determine a sequence of positive increments $\{\Delta_\ell\}$ with $\sum_{\ell=1}^{L} \Delta_\ell = 1$. Thus at the end of cycle ℓ, $k^{(\ell)}(\theta) = k(\theta)^{r_\ell}$ where $r_\ell = \sum_{s=1}^{\ell} \Delta_s$. This variant of the C phase is known as *power tempering* . The term originates in the simulated annealing literature in which $T_\ell = r_\ell^{-1}$ is known as *temperature* and $\{T_\ell\}$ as the *cooling schedule*. Another approach originates in particle filtering and Bayesian inference: $k^{(\ell)}(\theta) = p\left(y_{1:t_\ell} \mid \theta\right)$, where $0 < t_1 \ldots < t_L = T$ for a sample of size T. The increments are therefore $w^{(\ell)}(\theta) = p\left(y_{t_{\ell-1}+1:t_\ell} \mid y_{1:t_{\ell-1}}, \theta\right)$. This variant of the C phase is known as *data tempering*.

The C phase can be motivated informally by analogy to importance sampling, a long-established Monte Carlo simulation method, interpreting $k^{*(\ell-1)}(\theta)$ as the kernel of the source density and $k^{*(\ell)}(\theta)$ as the kernel of the target density. If it were the case that the particles $\theta_{jn}^{(\ell-1)}$ were independent and had common distribution indicated by the kernel density $k^{*(\ell-1)}(\theta)$, then

$$\begin{aligned} & \frac{\sum_{j=1}^{J}\sum_{n=1}^{N} w\left(\theta_{jn}^{(\ell-1)}\right) g\left(\theta_{jn}^{(\ell-1)}\right)}{\sum_{j=1}^{J}\sum_{n=1}^{N} w\left(\theta_{jn}^{(\ell-1)}\right)} \xrightarrow{a.s.} \frac{\int_{\Theta} k^{*(\ell)}(\theta)\, g(\theta)\, d\theta}{\int_{\Theta} k^{*(\ell)}(\theta)\, d\theta} \\ = & \int_{\Theta} p^{*(\ell)}(\theta)\, g(\theta)\, d\theta = E^{(\ell)}[g(\theta)] \end{aligned} \tag{12}$$

so long as $E^{(\ell)}[g(\theta)]$ exists. The convergence is in N, the number of particles per group.

The core of the argument for importance sampling is

$$\int_{\Theta} p^{*(\ell)}(\theta)\, g(\theta)\, d\theta = \frac{\int_{\Theta} w^{(\ell)}(\theta)\, k^{*(\ell-1)}(\theta)\, g(\theta)\, d\theta}{\int_{\Theta} w^{(\ell)}(\theta)\, k^{*(\ell-1)}(\theta)\, d\theta} = \frac{\int_{\Theta} w^{(\ell)}(\theta)\, p^{*(\ell-1)} g(\theta)\, d\theta}{\int_{\Theta} w^{(\ell)}(\theta)\, p^{*(\ell-1)}(\theta)\, d\theta}.$$

This result does not apply strictly, here, because while the particles $\theta_{jn}^{(\ell-1)}$ are identically distributed, they are not independent and $k^{*(\ell-1)}(\theta)$ is at best an approximation of the kernel density of the true common distribution of the particles $\theta_{jn}^{(\ell-1)}$ so long as $N < \infty$ (as it must be in practice). But many of the practical concerns in importance sampling carry over. In particular, success lies in $w(\theta)$ being "well-conditioned"–loosely speaking, variation in $w\left(\theta_{jn}\right)$ must not be too great. For example, difficulties arise when just a few weights $w\left(\theta_{jn}\right)$ account for most of the sum. In this case the target density kernel $k^{*(\ell)}(\theta)$ is represented almost entirely by a small number of particles and the approximation of $E^{(\ell)}[g(\theta)]$ implicit in the left side of (12) is poor.

The C phase directly confronts the key question of how much information to introduce in cycle ℓ: too little and L will be larger than it need be; too much, and it becomes difficult for the other phases to convert ill-weighted particles from cycle $\ell - 1$ into particles from cycle ℓ sufficiently independent that the representation of the distribution does not deteriorate from one cycle to the next into a state of gross unreliability. A conventional and effective way to monitor the quality of the weight function is by means of *relative effective sample size*

$$RESS^{(\ell)} = \frac{ESS^{(\ell)}}{JN} = \frac{\left[\sum_{j=1}^{J}\sum_{n=1}^{N} w^{(\ell)}\left(\theta_{jn}^{(\ell-1)}\right)\right]^2}{JN\sum_{j=1}^{J}\sum_{n=1}^{N} w^{(\ell)}\left(\theta_{jn}^{(\ell-1)}\right)^2}. \tag{13}$$

The *effective sample size* $ESS^{(\ell)}$ is an adjustment to the sample size (number of particles, JN) that accounts for lack of balance in the weights, and relative effective size is its ratio to sample size.

In general $RESS^{(\ell)}$ is lower the more information is introduced in the C phase. This is always true for power tempering and as a practical matter is nearly always the case for data tempering. It suggests a strategy of introducing no further information after $RESS^{(\ell)}$ has attained or fallen below a target value $RESS^*$. The target $RESS^* = 0.5$ is usually reasonable. Practical experience shows that somewhat higher $RESS^*$ leads to more cycles but faster execution in the M phase, lower $RESS^*$ to fewer cycles but slower M phase execution, and as a result there is not much difference in execution time over the interval $(0.1, 0.9)$ for $RESS^*$.

Before any new information is introduced in the C phase $w^{(\ell)}(\theta) = 1$. Data tempering entails iterations $s = 1, 2, \ldots$ in which iteration s introduces $y_{t_{\ell-1}+s}$, updates

$$w^{(\ell)}\left(\theta_{jn}^{(\ell-1)}\right) = w^{(\ell)}\left(\theta_{jn}^{(\ell-1)}\right) \cdot p\left(y_{t_{\ell-1}+s} \mid y_{t_{\ell-1}+s-1}, \theta_{jn}^{(\ell-1)}\right),$$

and computes the corresponding $RESS^{(\ell)}$. Iterations terminate the first time $RESS^{(\ell)} < RESS^*$. This procedure is well established, though much of the sequential Monte Carlo particle filtering literature introduces exactly one new observation per cycle. In neither approach does the algorithm control the amount of information introduced in the C phase and therefore in each cycle. Indeed, routine applications of SABL demonstrate that unusual or outlying observations (e.g., asset returns for days or periods marked by financial crisis) can produce $RESS^{(\ell)}$ with values much smaller than $RESS^*$, imply poor performance in the S and M phases, and generally compromise the efficiency of the algorithm.

Power tempering, discussed briefly at the start of this section, makes it possible in principle for the algorithm to control the introduction of information through the choice of the sequence of increments $\{\Delta_\ell\}$. Precisely the same problem arises in simulated annealing approaches to optimization, which uses temperature, the inverse of power, and the problem is known as choice of the temperature reduction schedule. We return to this problem in Section 5, because it has some closely related aspects that apply to optimization but not Bayesian inference. The solution developed there is subsequently used for both Bayesian inference and maximum likelihood estimation in Sections 4 and 6.

3.2. *S* phase

The rest of cycle ℓ starts with the weighted particles $\theta_{jn}^{(\ell-1)}$ from the end of the C phase and produces unweighted particles $\theta_{jn}^{(\ell)}$ that that meet or exceed a mixing condition—a measure of lack of dependence described in the next section—at the end of the M phase. The S phase begins this process, removing weights by means of resampling. The principle behind resampling is to regard the weight function as the kernel of a discrete probability function defined over the particles and draw from this distribution with replacement. Hence the name selection phase. SABL performs this operation on each group of particles separately—that is, particles are always selected within groups and never across groups. This independence between the groups $j = 1, \ldots, J$ is essential in (1) proving the convergence of the algorithm; (2) assessing the mixing condition in the M phase; and (3) providing a numerical standard error for the approximation as discussed in Section 3.4. Resampling produces unweighted particles denoted $\theta_{jn}^{(\ell,0)}$.

The most elementary resampling method is to make N independent and identically distributed draws from the multinomial distribution with argument N and probabilities

$$p_{jn} = w^{(\ell)}\left(\theta_{jn}^{(\ell-1)}\right) / \sum_{i=1}^{N} w^{(\ell)}\left(\theta_{ji}^{(\ell-1)}\right) \quad (n = 1, \ldots, N).$$

This method is known as *multinomial resampling*. An alternative method, known as *residual resampling*, is to compute the same probabilities and collect an initial subsample of size $N^* \leq N$ consisting of $\left[N \cdot p_{jn}\right]$ copies of each particle θ_{jn}, where the function $[\cdot]$ is standard notation for what is variously known as the greatest whole integer, greatest integer not greater than, or floor function. Then draw the remaining $N - N^*$ particles by means of multinomial resampling with probabilities $p_{JN}^* \propto Np_{jn} - \left[N \cdot p_{jn}\right]$. Residual resampling results in lower dependence amongst the particles $\theta_{jn}^{(\ell,0)}$ $(n = 1, \ldots, N)$ than does multinomial resampling. For both methods there are central limit theorems that are essential to demonstrating convergence and interpreting numerical standard errors. There are other resampling methods that lead to even less dependence amongst the particles, but for these methods central limit theorems do not apply. These methods are all described in [5].

The S phase is a simple but key part of the SABL algorithm. Resampling is also a key part of evolutionary (or, genetic) algorithms where it plays much the same role. The particles $\theta_{jn}^{(\ell,0)}$ $(n = 1, \ldots, N)$ are for this reason sometimes called the *children* of the *parent* particles $\left\{\theta_{jn}^{(\ell-1)}\right\}$ $(n = 1, \ldots, N)$, and also to emphasize the fact that for each child $\theta_{jn}^{(\ell,0)}$ there is a parent $\theta_{jn'}^{(\ell-1)}$. Parents with larger weights are likely to have more children—it is not hard to work out the exact distribution of the number of children of a given parent for any one parent for multinomial resampling and then again for residual resampling. With both, the expected number of children, or fertility, of the parent $\theta_{jn}^{(\ell-1)}$is proportional to $w\left(\theta_{jn}^{(\ell-1)}\right)$, a measure of the parent's "success" in the environment of the information introduced in cycle ℓ.

3.3. M phase

If the algorithm were to continue in this way, the number of unique children would never increase and in general would decrease from cycle to cycle. Indeed, in the context of Bayesian inference it can be shown under mild regularity conditions that the number of unique particles converges almost surely to 1 as the number of observations increases.

The M phase addresses this problem by creating diversity amongst sibling particles in a way that is faithful to the information kernel $k^{*(\ell)}(\theta)$. It does so using the same principle of invariance that is central to Markov chain Monte Carlo (MCMC) algorithms, drawing particles from a transition density $dQ^{(\ell)}(\theta \mid \theta^*)$ with the invariance property

$$\int_{\Theta} k^{*(\ell)}(\theta^*)\, dQ^{(\ell)}(\theta \mid \theta^*)\, d\theta^* = k^{*(\ell)}(\theta) \;\; \forall\, \theta \in \Theta. \tag{14}$$

The universe of invariant transition densities is large and manifest in the MCMC literature. Many of these transitions are model-specific, for example Gibbs sampling variants of MCMC. On the other hand a number of families of Metropolis-Hastings transitions apply quite generally and with problem-specific tuning of parameters can be computationally efficient.

SABL incorporates one of these variants, the Metropolis Gaussian random walk. The M phase applies the Metropolis random walk repeatedly in steps $s = 1, 2, \ldots$, each step generating a new set of particles $\theta_{jn}^{(\ell,s)}$ from the previous set $\theta_{jn}^{(\ell,s-1)}$. Following the familiar arithmetic, candidate new particles are generated $\theta_{jn}^{*(\ell,s)} \sim N\left(\theta_{jn}^{(\ell,s-1)}, \Sigma^{(\ell,s-1)}\right)$ and accepted with probability

$$\min\left[\frac{k^{*(\ell)}\left(\theta_{jn}^{*(\ell,s)}\right)}{k^{*(\ell)}\left(\theta_{jn}^{(\ell,s-1)}\right)},\ 1\right].$$

In SABL $\Sigma^{(\ell,s)}$ is proportional to the sample variance of $\theta_{jn}^{(\ell,0)}$ computed using all the particles. The factor of proportionality increases when the rate of candidate acceptance in the previous step exceeds a specified threshold and is decreased otherwise. This draws on established practice in MCMC and works well in this context. SABL incorporates variants of the basic Metropolis Gaussian random walk, as well, drawing on experience in the MCMC literature.

The objective of the M phase is to attain a degree of independence of the particles $\theta_{jn}^{(\ell)}$ at the end of each cycle sufficient to render the final set of particles $\theta_{jn} = \theta_{jn}^{(L)}$ a reliable representation of the distribution implied by the probability density function $p^*(\theta)$. The idea behind M phase termination in SABL is to measure the degree of mixing (lack of dependence) amongst the particles at the end of each Metropolis step s of cycle ℓ, and terminate when this measure meets or exceeds a certain threshold.

In SABL mixing is measured by the average relative numerical efficiency (RNE) of a group of functions chosen specifically for this purpose in each model. The RNE of the SABL approximation of a posterior moment $E[g(\theta)] = \int_\Theta g(\theta)\, p^*(\theta)\, d\theta$ is a measure of its numerical accuracy relative to that achieved by a hypothetical simulation $\theta_{ij} \overset{iid}{\sim} p^*(\theta)$. The next section explains how this measure is constructed. In the M phase the RNE of the particles $\left\{\theta^{(\ell,s)}\right\}$ tends to increase with the number of steps s, though not monotonically.

A simple stopping rule for the M phase is to terminate the iterations of the Metropolis random walk when the average RNE of a group of functions first exceeds a stated threshold. In any application there are practical limits to the average RNE that can be achieved through these iterations, and so SABL imposes a limit on their number. Achieving greater independence of particles is especially important in the last cycle, because at the end of the M phase in that cycle the particles constitute the representation of $p^*(\theta)$. The SABL core default criterion is average RNE 0.4 with 100 maximum iterations in cycles $1, \ldots, L-1$ and average RNE 0.9 with 300 maximum iterations in the final cycle L.

Mixing thoroughly is not the objective of the M phase. In MCMC that is essential in providing a workable representation of the distribution with kernel $k^*(\theta)$. In SABL the C and S phases take on this important task, whereas the function of the M phase is to place a lower bound on the dependence amongst particles.

3.4. Convergence and the Two-Pass Variant of SABL

Durham and Geweke [3] shows that bounded likelihood and existence of the relevant prior moment together sufficient for ergodicity. In all posterior simulators the assessment of numerical accuracy is based on a central limit theorem, which in this context takes the form

$$N^{1/2}\left(\overline{g}^{(J,N)} - \overline{g}\right) \xrightarrow{d} N\left(0, \sigma_g^2\right) \tag{15}$$

where

$$\overline{g} = \int_\Theta g(\theta)\, p^*(\theta)\, d\theta \quad \text{and} \quad \overline{g}^{(J,N)} = N^{-1}\sum_{n=1}^{N} g\left(\theta_{jn}\right).$$

By itself (15) is not enough: it is essential to compute or approximation σ_g^2 as well.

The theory developed in the sequential Monte Carlo literature provides a start. It posits a fixed pre-specified sequence of kernels $k^{(1)}, \ldots, k^{(L)}$ (see (11)) and a fixed pre-specified sequence of M phase transition densities $dQ^{(\ell)}$ (see (14)), together with side conditions (implied by bounded likelihood and the existence of prior moments), and proves (15). For example, this is the framework

set up in [8] as well as the careful treatment in [5]. But in any practical application the kernels $k^{(\ell)}$ and transition densities $dQ^{(\ell)}$ are adaptive, relying on information in the particles $\theta_{jn}^{(\ell-1)}$ or $\theta_{jn}^{(\ell,s-1)}$, rather than fixed. The theory does not apply then because the kernels and transitions depend on the random particles, and the structure of this dependence is so complex as to preclude extension of the existing theory to this case—especially for the transition kernels $dQ^{(\ell)}$. Thus, this literature provides a theory of sequential Bayesian learning but not a theory of *sequentially adaptive* Bayesian learning. It is universally recognized that some form of adaptation is required, for it is impossible to pre-specify kernels $k^{(\ell)}$ and transition densities $dQ^{(\ell)}$ that provide reliable approximations in tolerable time without knowing a great deal about the posterior distribution—which, of course, is the goal and not the starting point.

Durham and Geweke [3] deals with this issue by creating the two-pass variant of the algorithm. The first pass is exactly as described in this section, with the addition that the kernels $k^{(\ell)}$ and transitions $dQ^{(\ell)}$ are saved. For the specific variants described in Sections 3.1 and 3.3, this amounts to saving the sequence $\{r_\ell\}$ or $\{t_\ell\}$ from the C phase and the doubly-indexed sequence of variance matrices $\Sigma^{(\ell,s-1)}$ from the M phase, but the idea generalizes to other variants of the C and M phases. The second pass re-executes the algorithm (with different seeds for the random number generator) and uses the kernels $k^{(\ell)}$ and transitions $dQ^{(\ell)}$ computed in the first pass, skipping the work required to compute these objects from the particles. The theory developed in the sequential Monte Carlo literature then applies directly to the second pass, because the kernels $k^{(\ell)}$ and transitions $dQ^{(\ell)}$ are in fact fixed in the second pass.

Experience thus far is that substantial differences between the first and second passes do not arise, and can only be made to do so by specifying imprudently small values of N. Thus in practice it suffices to use the two-pass algorithm only occasionally—perhaps at the inception of a research project when the general character of the model(s), data and sample size are known, and then again prior to communicating findings.

The sequential Monte Carlo literature provides abstract expressions for σ_g^2 in (15) but no means of evaluating or approximating σ_g^2. SABL provides the approximation using the particle groups. Consider the second pass of the two-pass algorithm where the convergence theory fully applies. In this setting there is no dependence of particles across groups. The M phase and the C phase are perfectly parallel: exactly the same operations applied to all the particles with no communication between particles. Resampling in the S phase, which introduces dependence amongst particles, takes place entirely within groups so as not to compromise independence across groups. Therefore the approximations $\overline{g}_{jN} = N^{-1}\sum_{n=1}^{N} g\left(\theta_{jn}\right)$ of $\overline{g} = E\left[g\left(\theta\right)\right]$ are independent across the groups $j = 1, \ldots, J$. A central limit theorem (15) applies within each group so long as $g\left(\theta\right)$ has finite second moment. Computing the cross-group mean $\overline{g}_{J,N} = J^{-1}\sum_{j=1}^{J}\overline{g}_{jN}$, a conventional estimate of σ_g^2 in (15) is

$$\widehat{\sigma}_g^2 = N \cdot (J-1)^{-1} \sum_{j=1}^{J} \left(\overline{g}_{jN} - \overline{g}_{J,N}\right)^2 \tag{16}$$

and

$$(J-1)\,\widehat{\sigma}_g^2/\sigma_g^2 \xrightarrow{d} \chi^2\,(J-1)\,, \tag{17}$$

the convergence in (17) being in particles per group N. In the limit $N \to \infty$, $\overline{g}_{J,N}$ and $\widehat{\sigma}_g^2$ are independent.

The corresponding numerical variance estimate for $\overline{g}_{J,N}$ is

$$\widehat{\sigma}_{g,JN}^2 = (JN)^{-1}\,\widehat{\sigma}_g^2. \tag{18}$$

This should not be confused with the approximation of the posterior variance $\widehat{\mathrm{var}}\,(g) = (JN)^{-1}\sum_{j=1}^{J}\sum_{n=1}^{N}\left[g\left(\theta_{jn}\right) - \overline{g}_{J,N}\right]^2$. The *numerical standard error* corresponding to (18) is $\widehat{\sigma}_{g,JN} =$

$\left[\widehat{\sigma}^2_{g,JN}\right]^{1/2}$. This is the measure of accuracy used in SABL. From (17) the formal interpretation of numerical standard error is $\left(\overline{g}_{J,N} - \overline{g}\right) / \widehat{\sigma}_{g,JN} \xrightarrow{d} t\,(J-1)$. If particles within groups are independent then $\widehat{\sigma}^2_g \cong \widehat{\text{var}}\,(g)$, whereas if they are not then usually $\widehat{\sigma}^2_g > \widehat{\text{var}}\,(g)$, although $\widehat{\sigma}^2_g < \widehat{\text{var}}\,(g)$ may occur and is more likely with smaller numbers of particle groups J. The *relative numerical efficiency* of the approximation $\overline{g}_{J,N}$ is

$$RNE_g = \widehat{\text{var}}\,(g) / \widehat{\sigma}^2_g. \tag{19}$$

A useful interpretation of (19) is that a hypothetical simulator with $\theta_{jn} \overset{iid}{\sim} p^*(\theta)$ would achieve the same accuracy with $RNE_g \cdot JN$ particles.

This argument does not apply directly in the first pass because of the adaptation. In particular, recall that RNE is used in the M phase to assess mixing and determine the end of the sequence of iterations of the Metropolis random walk. This is an example of the complex feedback between particles and adaptation in the algorithm that has frustrated central limit theorems. This shortfall in theory is likely to persist. The two-pass procedure overcomes the problem and, moreover, provides the foundation for future variants of the algorithm without the overhead of establishing convergence for each variant.

4. Posterior Distributions of Half-Lives, Periods and Shocks

With the SABL infrastructure in place drawing a posterior sample is very simple and extremely fast.

4.1. Priors and Data

The prior distributions used in the quantitative results presented here are

Parameter	Distribution	Centered 90% interval
Intercept β_0 (1):	$\beta_0 \sim N\left(10, 5^2\right)$	$\beta_0 \in (1.77, 18.22)$
Secular half-life h_s:	$\log(h_s) \sim N(\log(25), 1)$	$h_s \in (5.591, 111.9)$
Cyclical half-life h_c:	$\log(h_c) \sim N(\log(1), 1)$	$h_c \in (0.2237, 22.36)$
Period p:	$\log(p) \sim N(\log(5), 1)$, $p > 2$	$p \in (2.316, 28.46)$
Shock σ (1)	$\log(\sigma) \sim N(\log 0.025, 1)$	$\sigma \in (0.005591, 0.1118)$

All the priors but the first are substantive, that is, grounded in consideration of what is reasonable and what is not. The standard deviation 1 for these four corresponds to change in the parameter (e.g., h_s) by a factor of $e = 2.718$.

Given the interpretation of the parameters, I think it unlikely that an economist would claim as reasonable any values outside the stated 90% intervals. Indeed, many would be comfortable with more concentrated prior distributions. As will be seen, while the data contribute more to the posterior distributions than do the priors, the contribution of priors for half-lives and period is substantial. The effects of changing the prior hyperparameters are not hard to assess because all five parameters are nearly independent in the posterior distribution as well as in the prior.

The data used in this exercise are annual OECD per capital real GDP constant purchasing power parity [9]. Data for 26 countries are available from 1970 through 2014, so given the three lags in (1) there are 42 annual observations. We concentrate on a more detailed presentation for the US, UK and Japan rather than in drawing comparisons and conclusions for all countries.

4.2. Computation

All of the results presented here were obtained using SABL Edition 2015a [1], a Matlab toolbox that has many options including massively parallel execution on graphics processing units as well as conventional muticore CPU execution. Some options important to the work here are

1. Choice of Bayesian or maximum likelihood inference, depending on a single parameter setting;
2. A system for mapping fundamental parameters (here, the vector θ specified in (2)) to the parameters of the normal model (here, β and σ in (1)) using a generic mapping system that applies in all models (here, the map is given by (2)–(4));
3. A system for specifying different prior distributions, either conventional or customized, for different components. In most cases there are options for truncation of prior distributions, like that for $\log(p)$.

Using SABL entails writing a single Matlab function for which there are templates included in the toolbox.

Table 1. SABL output, US posterior distribution.

```
SABL normal model
SABL executing using CPU with 1 worker
C phase anneal_Bayes algorithm
Effective sample size criterion (C.Cstop.ress) = 0.500
C phase Cstop_unconditional algorithm
S phase residual resampling
M phase MGRW_simple algorithm
M phase Mstop_rne stopping algorithm

Simulating 16384 times from truncated prior...  2.5324 seconds
Cycle 1 Cphase:  Likelihood function exponent 8.5448e-03, RESS 0.5000
Cycle 1 Sphase:  8565 particles out of 16384 unique (0.5228)
Cycle 1 Mphase:  1 iterations, mean RNE = 0.8175
Cycle 2 Cphase:  Likelihood function exponent 7.1163e-02, RESS 0.5000
Cycle 2 Sphase:  7533 particles out of 16384 unique (0.4598)
Cycle 2 Mphase:  5 iterations, mean RNE = 0.4875
Cycle 3 Cphase:  Likelihood function exponent 2.0704e-01, RESS 0.5000
Cycle 3 Sphase:  8374 particles out of 16384 unique (0.5111)
Cycle 3 Mphase:  1 iterations, mean RNE = 0.4208
Cycle 4 Cphase:  Likelihood function exponent 4.9638e-01, RESS 0.5000
Cycle 4 Sphase:  5084 particles out of 16384 unique (0.3103)
Cycle 4 Mphase:  4 iterations, mean RNE = 0.4472
Cycle 5 Cphase:  Likelihood function exponent 1.0000e+00, RESS 0.6051
Cycle 5 Sphase:  7471 particles out of 16384 unique (0.4560)
Cycle 5 Mphase:  14 iterations, mean RNE = 0.9179
Elapsed clock time 5.52 seconds
   CPU time 7.71 seconds
   Ratio 1.40
```

Table 1 provides the output from the computation of the posterior distribution for the US data. It uses the same terminology as Section 3. All of the technical parameters associated with the SABL algorithm are the default values; any of them can be changed in the single function the user writes to interface with SABL. Execution with the default number of particle groups $(J = 8)$ and particles per group $(N = 1024)$ required less than 6 seconds on a standard laptop with a quadcore CPU. Of this, almost half the time was taken up generating the initial values from the truncated prior distribution for $\log(p)$. Notice that the relative effective sample size is exactly the target value in each cycle. Each M phase continues until RNE 0.4 or greater has been achieved, but the last cycle applies the higher standard of 0.9 in order to provide more information in the final set of particles. Thus the 16,384 particles are nearly independent. It is easy to access any posterior moment along with its numerical standard error. Log marginal likelihood is produced as a by-product.

4.3. Findings

Tables 2–4 provide the first two posterior moments of the functions of interest $\log(h_z)$, $\log(h_c)$, $\log(p)$ and $\log(\sigma)$. Two aspects of these results are striking. First, the results for the three countries are quite similar. Given the integration of the global economy in this period and the contemporaneous data, this is perhaps not surprising. Second, in each case the four parameters are nearly uncorrelated in the posterior distribution. This is in marked contrast to the situation for the parameter vector β in (1), where there is large correlation due the multicolinearity in $(y_{t-1}, y_{-2}, y_{t-3})$. This supports the efficacy of working with the chosen parameterization.

Table 2. US posterior moments.

Parameter		Mean	st. dev.
Secular $\log(h_s)$		4.056	0.667
Cyclical $\log(h_c)$		−0.584	0.548
Period $\log(p)$		2.045	0.565
Shock $\log(\sigma)$		−3.916	0.113
Correlation matrix:			
1.000	0.010	−0.021	0.001
0.101	1.000	−0.366	−0.018
−0.021	−0.366	1.000	0.022
0.001	−0.018	0.022	1.000

Table 3. UK posterior moments.

Parameter		Mean	st. dev.
Secular $\log(h_s)$		4.002	0.655
Cyclical $\log(h_c)$		−0.292	0.479
Period $\log(p)$		2.041	0.502
Shock $\log(\sigma)$		−3.885	0.106
Correlation matrix:			
1.000	0.027	−0.034	0.022
0.027	1.000	−0.515	−0.060
−0.034	−0.515	1.000	0.078
0.022	−0.060	0.078	1.000

Table 4. Japan posterior moments.

Parameter		Mean	st. dev.
Secular $\log(h_s)$		3.470	0.654
Cyclical $\log(h_c)$		−0.951	0.543
Period $\log(p)$		2.101	0.641
Shock $\log(\sigma)$		−3.806	0.111
Correlation matrix:			
1.000	0.030	0.084	0.133
0.030	1.000	−0.200	0.041
0.084	−0.200	1.000	−0.003
0.133	0.041	−0.003	1.000

Figures 1–3 provide the posterior density for each parameter, computed using a standard kernel smoothing algorithm optimized for normal distributions, together with the normal prior distribution for each. (The truncation for the $\log(p)$ prior is not shown.) Since all four parameters are transformed to logarithms, a common measure of the information in prior plus data is available. From Tables 2–4

and the northwest panel in each figure, over a 90% posterior credible interval secular half-life h_s changes by a factor of about 12 to 15. For h_c and p the range is about 8 to 9, and the factor is about 8 to 9, and for σ it is about 1.5 to 1.6. The relative confidence is not surprising: in a sample of any given length there are fewer secular fluctuations (which take many years) than there are cyclical fluctuations (which take a few years), whereas there is almost no dependence across observations in the information provided for σ.

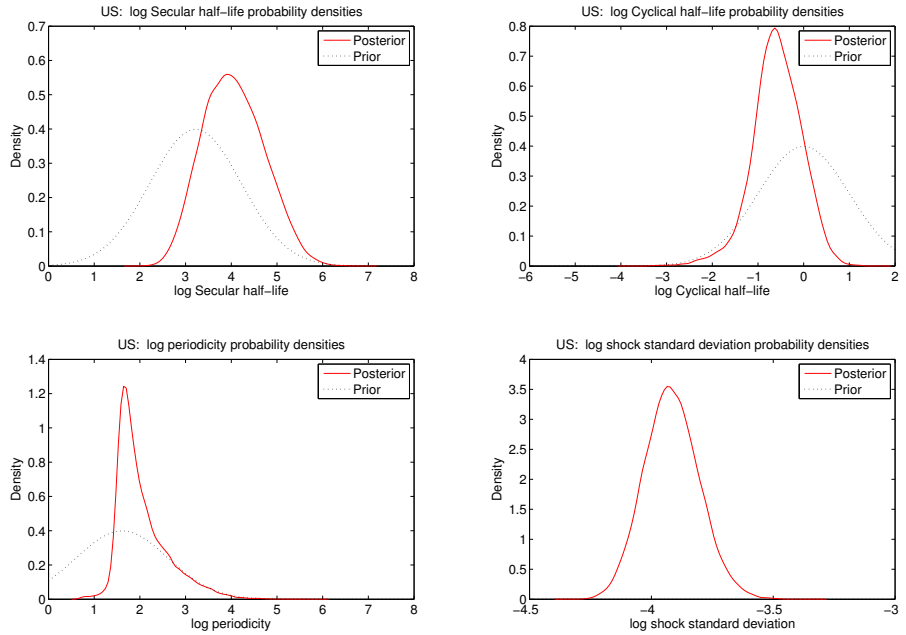

Figure 1. Prior and posterior parameter distributions, US.

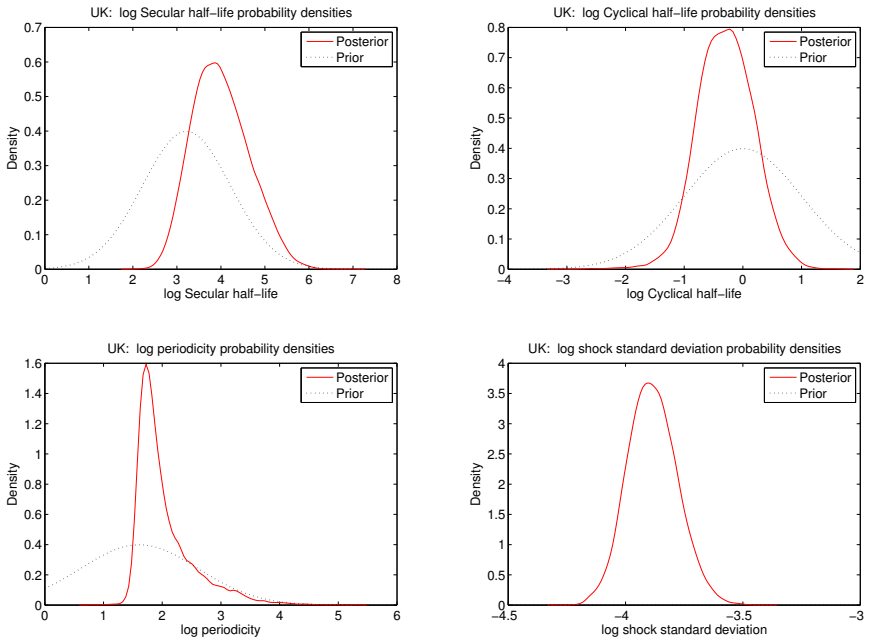

Figure 2. Prior and posterior parameter distributions, UK.

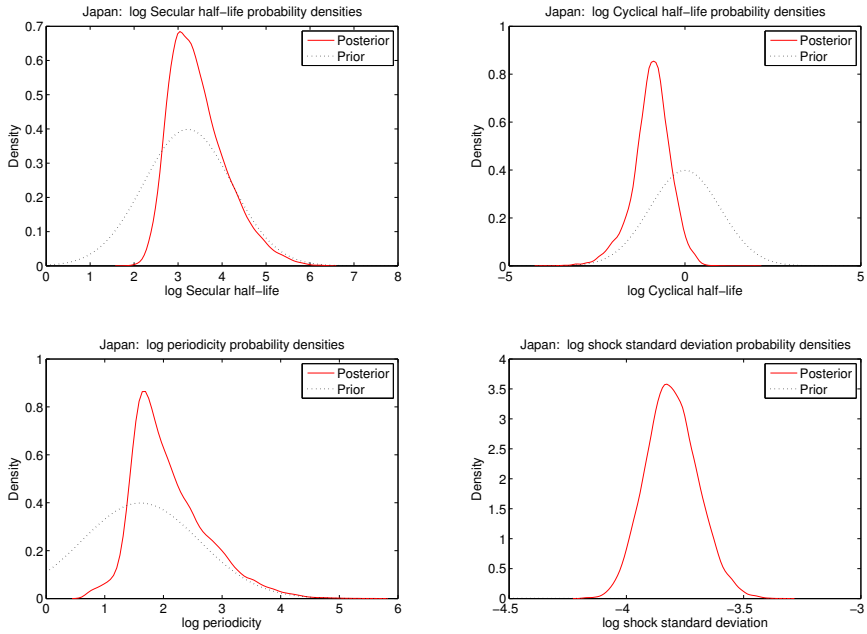

Figure 3. Prior and posterior parameter distributions, Japan.

The mechanics of how the prior distribution influences the posterior are straightforward in this situation. Since there is no correlation between parameters in the data and little in the posterior, by implication prior and data combine for each parameter with little interaction. Given the standard deviations of the prior distribution and the posterior distribution (Tables 2–4), data precision is only modestly higher than prior precision. Thus a location shift of a prior distribution by v units will produce a shift of a little less than $v/2$ units in the mean of the posterior distribution. Section 6 will compare these posterior distributions with those that would be obtained using the "diffuse" prior distribution in [7].

5. Maximum Likelihood Using SABL

With minor modification of the C phase, SABL handles global optimization problems as well as inference problems. This section provides a heuristic approach; for more formal motivation and technical details, see [2] and [4].

5.1. The Method

To begin, consider replacing the kernel $k(\theta)$ of Section 3 with the kernel $k(\theta)^q$ with $q > 1$. The corresponding probability density is now more concentrated than it was originally. The Bayesian annealing algorithm could proceed in precisely the same way but now terminating with $r_L = q$ rather than $r_L = 1$. Of course, the particles at that point would not correspond to a posterior distribution: such an interpretation would amount to an erroneous replication of the sample and additional $q - 1$ times. But everything stated in Section 3 about approximation of the distribution with kernel $k(\theta)$ remains true for the distribution with kernel $k(\theta)^q$.

Next consider what happens as q moves to increasingly higher values, and to this end two properties of $k(\theta)$ are useful:

1. The kernel $k(\theta)$ has a unique global mode θ^*;
2. $\log k(\theta)$ is twice continuously differentiable with bounded third derivative in a neighborhood of θ^*, and $\partial^2 \log k(\theta) / \partial\theta\partial\theta' \mid_{\theta=\theta^*} = H$, a negative definite matrix.

In the applications in this paper $k(\theta)$ is the likelihood function and similar conditions are invoked in standard limit theorems for posterior distributions and maximum likelihood estimators. The difference is that here the limit is $q \to \infty$ rather than increasing sample size.

Recall that in cycle ℓ of the annealing variant of the SABL algorithm the exponent of $k(\theta)$ becomes r_ℓ. Denote the variance matrix of particles at the end of this cycle by V_ℓ.

The following four implications follow, the first from the first property and the others from both. The first three are unsurprising given conventional results for the limits of posterior distributions and maximum likelihood estimators.

1. The probability distribution corresponding to the kernel $k(\theta)^q$ converges in distribution to the point θ^*.
2. The probability distribution of $q^{1/2}(\theta - \theta^*)$ converges in distribution $N\left(0, H^{-1}\right)$.
3. Taking the limit first as the number of particles N increases and then as the cycle ℓ increases,

$$r_\ell^{-1} V_\ell \xrightarrow{a.s.} H^{-1}, \tag{20}$$

   which is the asymptotic variance of the maximum likelihood estimator.
4. Taking limits in the same way, the power increase ratio $\rho_\ell = (r_\ell - r_{\ell-1}) / r_{\ell-1}$ converges almost surely to

$$\rho = RESS^{*-2/d} - 1 + \left[\left(RESS^{*-2/d} - 1\right) \cdot RESS^{*-2/d}\right]^{1/2}, \tag{21}$$

   where d is the dimension of θ. Note that ρ, the asymptotic power increase ratio, depends on $k(\theta)$ only through d.

A strength of this approach is that there is no need to determine first and second derivatives, either analytically or by means of numerical differentiation. The only requirements are (a) verify properties 1 and 2 analytically, (b) derive the likelihood function, (c) code the evaluation of the log-liklelihood function, and (d) code nonlinear parameter transformations. For many applications these requirements become trivial if one is using a nonlinear parameterization of an existing model, thus avoiding (b) and (c). In particular this approach avoids analytical derivation and code testing for first or second derivatives, or numerical evaluation of derivatives, both of which can be time consuming, tedious and encounter arcane numerical problems.

The results are insensitive to the prior distribution so long as the prior distribution provides support in a neighborhood of θ^*, the same condition invoked in the derivation of the asymptotic properties of posterior distributions. A practical consideration, here, is that as prior probability in this neighborhood decreases, r_1 becomes smaller and the number of cycles required to achieve a given concentration of particles becomes greater.

5.2. Convergence Criteria

Important practical questions are SABL termination criteria and the cycle(s) ℓ whose particles are used to approximate $V = H^{-1}$ using V_ℓ. Clearly the cycles chosen to approximate v must be cycles after the asymptotic power increase ratio has been closely attained, evidenced by fluctuations above and below ρ over a sequence of successive cycles. One may also wish to apply criteria for concentration of the distribution of the particles θ_{jn} or the distribution of the objective $\log\left[k\left(\theta_{jn}\right)\right]$. There seems no reason to continue beyond these points.

In fact, if cycles are allowed to continue, then eventually the algorithm encounters the limits of 64-bit arithmetic in distinguishing between values of $\log k\left(\theta_{jn}\right)$. The telltale evidence of this condition is that the evaluation of $\log k\left(\theta_{jn}\right)$ at different particles θ_{jn} reflects the bits of the mantissa corresponding to lowest significance. Experience with examples like the one in this paper strongly suggests that the sequence of power increase ratios $\{\rho_\ell\}$ exhibits three episodes, going to the limits of 64-bit arithmetic.

1. In the early cycles ρ_ℓ differs substantially from ρ. The most common pattern observed is that in the early cycles ρ_ℓ exceeds ρ, drifting downward toward ρ.
2. In the middle cycles ρ_ℓ fluctuates around ρ, fluctuations being smaller the larger the number of particles. With the default number of particles in the SABL software (2^{14}) fluctuations are less than 10% and commonly less than 5%. In these cycles the particles $\theta^{(\ell)}$ have a distribution that is hard to distinguish from multivariate normal.
3. In the later cycles ρ_ℓ drops well below ρ and fluctuates erratically. The reason is that variation in the particles is no longer dominated by the asymptotics outlined above: the limits of machine precision become increasingly important. (In the limit the distribution can become bizarre, as detailed in [2]). The random walk Gaussian Metropolis steps in the *M* phase are poorly suited to the the objective function now dominated by lower-order bit arithmetic, relative numerical efficiency is poor, and the particle distribution provides a poor source distribution in the *C* phase, leading to smaller increases in power in each cycle.

These patterns are generally evident in a plot of $\log \rho_\ell$ as a function of cycles ℓ, in which the middle cycles exhibit as a nearly flat portion of an otherwise generally decreasing function of ℓ. Any one of the middle cycles is a natural point to harvest the asymptotic variance matrix of the maximum likelihood estimator, as described above. At this point it is also natural to take as the MLE $\widehat{\theta} = \arg\max_{\ell,j,n} \log k\left(\theta_{jn}^{(\ell)}\right)$. Experience suggests that the accuracy of the MLE, so computed, is well beyond the number of digits typically reported.

5.3. Relationship to the Simulated Annealing Literature

The technique of simulated annealing was introduced over 30 years ago [10] and is now widely applied in science and engineering. Applications in statistics and econometrics are very rare and most econometricians do not include it in their suite of approaches to maximization problems. This is so despite the fact that [11] demonstrated the method's utility in that capacity in a *Journal of Econometrics* paper that ranks 18th (out of thousands) by citation in the simulated annealing literature, and ninth in the citation rankings of articles in that journal. Almost all of the citations come from the science and engineering literature, very few from economics or statistics.

The simulated annealing literature also uses a sequence of increasing powers of the objective function, but casts the sequence as its inverse r_ℓ^{-1}, known as *temperature*. Thus temperature decreases as the algorithm proceeds and this is responsible for its name. In the original version, and indeed most applications since, it may be regarded as a simple version of the SABL algorithm in which there is one particle and no *S* phase. Since there is no distribution of particles, neither the sequence of increasing powers in the *C* phase nor the variance in the Metropolis steps of the *M* phase can be constructed algorithmically as they are in SABL. Instead they must be provided directly by the user in each application.

Choosing a sequence of increasing powers (decreasing temperatures) that results in reasonable efficiency—or even works at all—has been a challenge. Typical applications, even by experienced practitioners, involve trial, error and tinkering with temperature reduction schedules. The choice of the variance matrix for the Metropolis step similarly must be tailored to each application, a procedure familiar to many Bayesian econometricians and statisticians who have used the Metropolis random walk for Bayesian inference. In simulated annealing, this variance matrix must be related systematically to temperature (power).

Very recently the simulated annealing literature has begun to adopt parallel chains of arguments–particles, in the terminology of this paper. Zhou and Chen [12] is representative. I am not aware of any work in this literature that attempts to use the parallel chains to address the temperature reduction problem and the Metropolis variance matrix problem systematically. Thus the method still suffers from this very substantial overhead in application. Once acceptable solutions to these problems have been found by trial and error, none begins to approach the standard of computing the maximizing argument to the limits of machine precision attained by SABL. Geweke

and Frischknecht [2] draws an explicit comparison using the test problems in [12], showing that SABL achieves machine precision with about the same number of floating point operations used in the methods of that paper, which in fact do not even determine the optimizing values to even three significant figures in all cases.

6. Maximum Likelihood Estimation of Half-Lives and Periods

This section illustrates optimization in SABL for the case of maximum likelihood using the U.S. data. As with the posterior distribution results are quite similar across countries.

6.1. Performance of the Optimization Algorithm

Figure 4 provides a global perspective on the behavior of the algorithm. The northwest panel provides the power r (of the likelihood function in each cycle of the SABL algorithm), and the south east panel displays the power increase ratio. In the terminology of the simulated annealing literature the temperature is the inverse of power, and the analogue of temperature for the northwest panel would simply flip the graph about a horizontal axis of rotation at 10^0. The temperature decay and power increase ratios are identical. In both cases it is straightforward to identify the cycles over which the geometric rate of increase is constant, roughly cycles 14 through 46. The power increase (temperature decay) ratio fluctuates about the theoretical value ρ stated in (21) and shown by the dotted line in the southeast panel. The last cycle in which the ratio exceeds ρ is $\ell = 44$. Computation to this point required 44 seconds using the same hardware and software used for the posterior distribution in Section 4. The difference is due to the fact that Bayesian inference required only 5 cycles (Table 1).

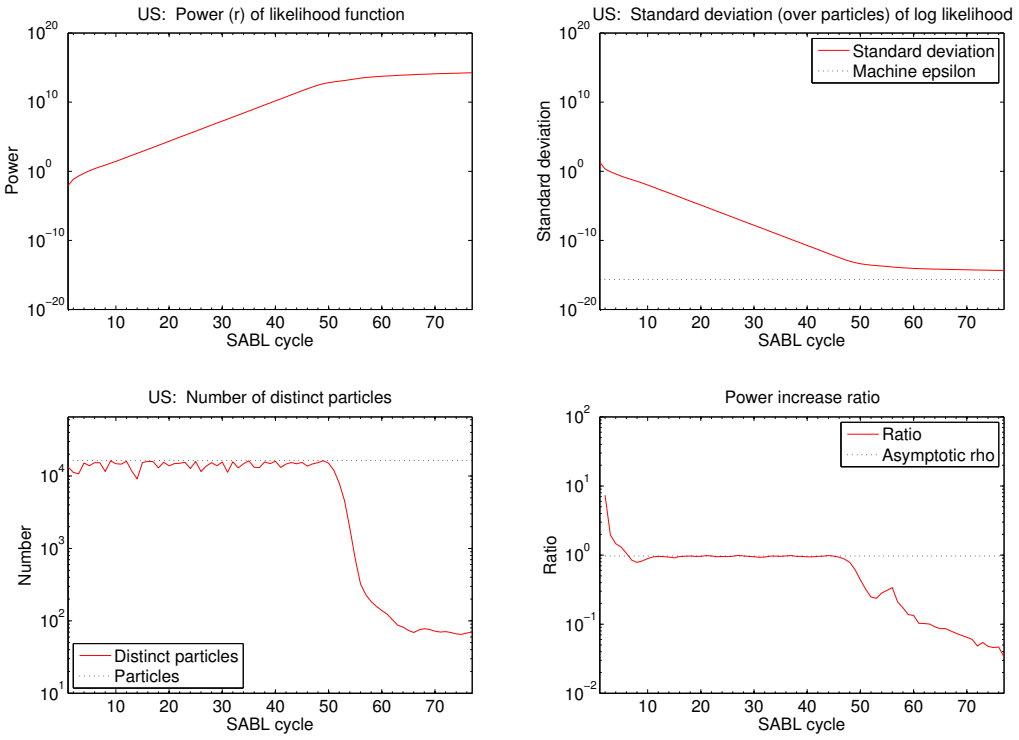

Figure 4. Convergence details for the SABL optimization algorithm.

The cycles of constant power increase ratio end when the limitations of floating point arithmetic begin to disguise the true quadratic nature of the log-likelihood function in the neighborhood of the maximum. Up to this point the Metropolis random walk is effective in mixing particles in the M phase, which may be seen in the fact that at most 8 iteration are required to attain the RNE criterion of

0.4 that ends the M phase and the cycle, whereas from iteration 49 onward the convergence criterion is never met and the M phase goes to the default limit of 100 iterations. After iteration 48 (roughly) RNE drops rapidly, the weight function in the C phase is poor, and the number of distinct particles (southwest panel) deteriorates quickly. By iteration 60 there are only about 100 distinct particles.

The northeast panel provides a complimentary perspective on the limitations imposed by machine arithmetic. So long as the power increase ratio is near ρ, the standard deviation of the log-likelihood objective function is very nearly inversely proportional to power. In all events differences between real numbers are multiples of what is known as "machine epsilon", which for conventional 64-bit floating point representation is $\varepsilon = 2.22 \times 10^{-16}$. This graininess becomes a factor beyond about iteration 48. Standard deviation of the objective function, across particles, decreases toward ε, shown by the dotted line in the northeast panel. Eventually the objective function behaves like a step function and except for extremely simply objective functions the steps appear random to the eye.

Figure 5 provides perspectives on the maximum likelihood estimates and asymptotic standard errors in each cycle of the SABL algorithm. The maximum likelihood estimates in the upper panel are constant (to the accuracy of the plot) beyond about cycle 20. The complications of machine arithmetic affect only the evaluated shape of the surface, not its location, and so these estimates remain unaffected. The deduction of asymptotic standard errors from the distribution of the particles, on the other hand, is closely tied to multivariate normal distribution of particles. It is constant over the same range that the power increase ratio is constant, and exhibits about the same amplitude of relative fluctuations. Beyond about cycle 54 it increases, due to the fact that the limitations of 64-bit arithmetic introduce variation in particles that is significant relative to the actual value, meaning that it continues to increase. The lower panel of Figure 5 supports the convergence criterion of last cycle $(\ell = 46)$ in which the power decay ratio exceeds ρ.

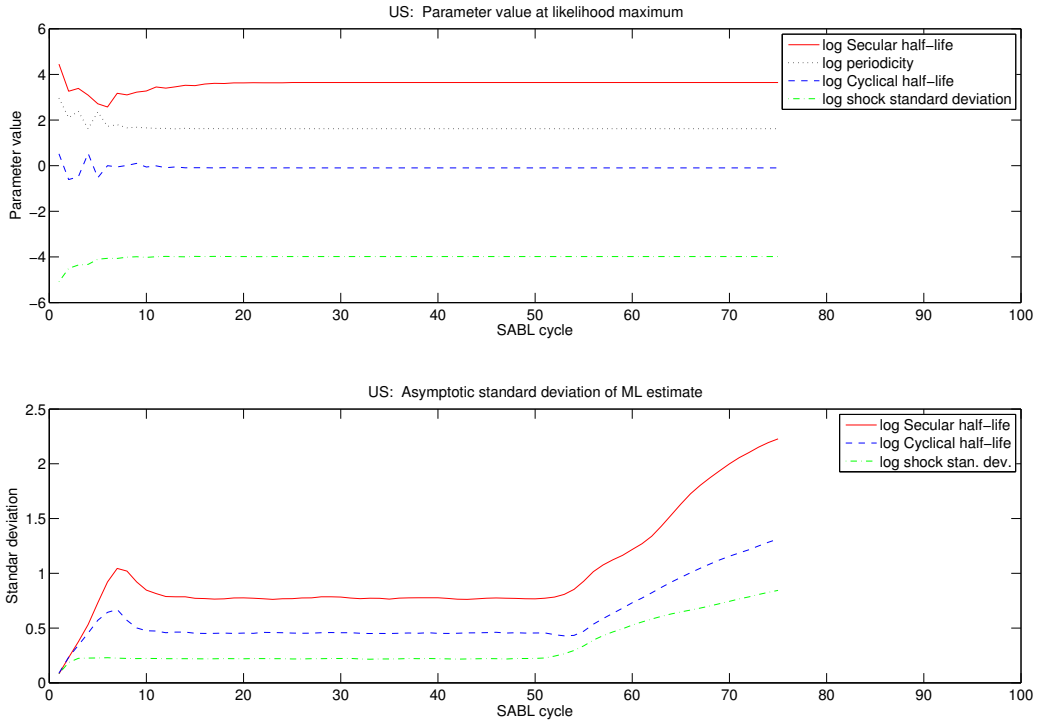

Figure 5. Implied maximum likelihood estimates and asymptotic standard errors by cycle.

Figure 6 supplements these perspectives by providing the distribution of the parameter $\theta_4 = \log(p)$ in some cycles of interest. (Result for other parameters are qualitatively similar.) The solid line provides the kernel smoothed estimate of the distribution of particles, the dotted line

the Gaussian distribution corresponding to particle mean and variance. In all cases the deviation of values from the final maximum likelihood estimate are shown to make the horizontal axis ticks readily interpretable. Prior to the cycles of constant power increase ratio the shape of the distribution moves from the posterior distribution portrayed in Figure 1 (cycle 5 is very close) to the quadratic expansion of the log-likelihood about the maximum likelihood estimate, scaled by power. Beyond that point the distribution becomes erratic as it is increasingly corrupted by the limitations of machine arithmetic. The Metropolis step, with its Gaussian proposal is increasingly ineffective; to appreciate this fully, recall that the Metropolis step is dealing with this behavior in five dimensions at once.

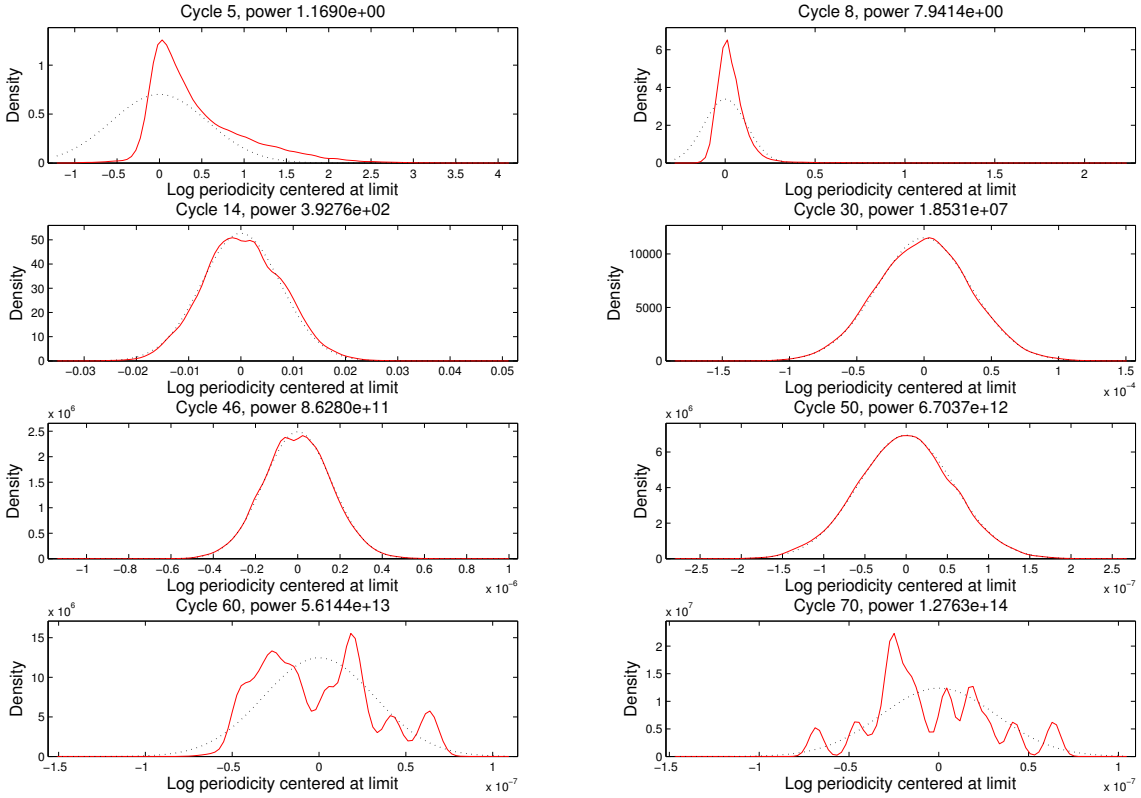

Figure 6. Distribution of particles at selected cycles, SABL optimization algorithm.

These results support the idea that iteration to maximum likelihood can halt when the power increase ratio begins to fall from the neighborhood of the value ρ implied by a quadratic objective function of the relevant dimension. At the last cycle in which this ratio exceeds ρ , the maximum likelihood estimate can be taken to be the particle providing the highest log-likelihood and the asymptotic variance can be taken to be particle variance multiplied by power r_ℓ.

6.2. Findings

It is straightforward to compute maximum likelihood estimates and their asymptotic variance as just described. We compare these results with two other approaches. One is the posterior distribution presented in Section 4. The other begins with the closed-form maximum likelihood (least squares) estimate of (β, σ) in (1) and transforms the estimate to $(\beta_0, \log(h_s), \log(h_c), \log(p), \log(\sigma))$ as described in that section. By the invariance property of maximum likelihood estimates, these should be the same as those obtained using the methods of Section 5. From the last procedure we also compute the distribution of $(\beta_0, \log(h_s), \log(h_c), \log(p), \log(\sigma))$ obtained by mapping from the asymptotic distribution of the maximum likelihood estimator of (β, σ^2). This will not be the same as asymptotic distribution the direct maximum likelihood estimates. But it comes very close to the posterior distribution in [7] because the prior distribution in that work was $p(\beta.\sigma) \propto \sigma^{-1}$ subject

to the constraints of stationarity and two complex roots of the lag operator generating polynomial. (Those constraints are satisfied in over 99% of the draws from the asymptotic normal distribution of (β, σ) here.) Thus, this comparison enables us to examining the influence of two quite different prior distributions for $(\beta_0, \log(h_s), \log(h_c), \log(p), \log(\sigma))$.

Table 5 provides the first two moments of each parameter for the three cases. Those for the posterior distribution are the same as in Table 2. The maximum likelihood estimates of $\log(h_c)$ and $\log(p)$ differ from the posterior means, by more than one standard error in the first case and by over two maximum likelihood standard errors in the second (using the metric of the direct ML standard error in each case). Except for $\log(p)$ the posterior standard deviation and the ML asymptotic standard errors are similar. For $\log(p)$ the direct ML asymptotic standard error is smaller than the posterior standard deviation by a factor of almost 4 and the indirect ML asymptotic standard error by a factor of over 2. This gross divergence for $\log(p)$ is consistent with the posterior density in the southwest panel of Figure 1, whose model is much more sharply defined than is the mode of a normal distribution. The conventional local expansion of the log-likelihood function is unrepresentative of its global behavior in the case of $\log(p)$.

Table 5. US posterior moments and maximum likelihood estimates.

	Posterior		**Direct ML, θ**		**OLS Indirect ML**	
Parameter	mean	st. dev.	MLE	st. dev.	MLE	st. dev.
Secular $\log(h_s)$	4.056	0.667	3.646	0.767	3.646	0.866
Cyclical $\log(h_c)$	−0.584	0.548	−0.096	0.457	−0.096	0.406
Period $\log(p)$	2.045	0.565	1.621	0.148	1.621	0.231
Shock $\log(\sigma)$	−3.917	0.112	−3.984	0.109	−3.984	0.117

Figure 7 provides another perspective on the comparison between the three approaches to inference, for the joint distribution of secular and cyclical half-life (left panels) and the joint distribution of period and cyclical half-life (right panels). To facilitate comparison the axes are identical in each column and were selected so as to exclude the 0.25% smallest and 0.25% largest points out of 820 selected from the 16,348 particles in the posterior distribution. The number of points plotted in the last two rows is almost 820 since the dispersions there are smaller. The horizontal and vertical lines are the same in each column, intersecting at the maximum likelihood estimate, which is indicated by the circle in the last two rows. In the top pair of panels the circle is the mean of the posterior distribution and the cross is the median. The smaller dispersion of the parameters under the asymptotic ML distributions, especially for $\log(p)$, is striking in these figures. So, too, are the differences in shape. The joint distribution under the direct ML asymptotic expansion is Gaussian by construction; as already documented and well-understood, the posterior distribution is not; but the indirect ML distribution is also non-Gaussian because the transformation from (β, σ) to θ is nonlinear.

The indirect ML asymptotic expansion is very close to the exact posterior distribution using the conventional improper prior distribution $p(\beta.\sigma) \propto \sigma^{-1}$, sometimes called the Jeffreys prior on (less precisely) "diffuse" or "uninformative."uninformative prior. This is in fact not a Jeffreys prior for linear regression except when it reduces to mean estimation. Whereas the prior in β is flat, the implied prior for θ is not due to the Jacobian implied by the nonlinear transformation in Equations (2) through (4). The fact that "diffuse "or "uninformative"prior distribution can be vague, slippery or vacuous concepts is sometimes missed by Bayesian econometricians, like [7].

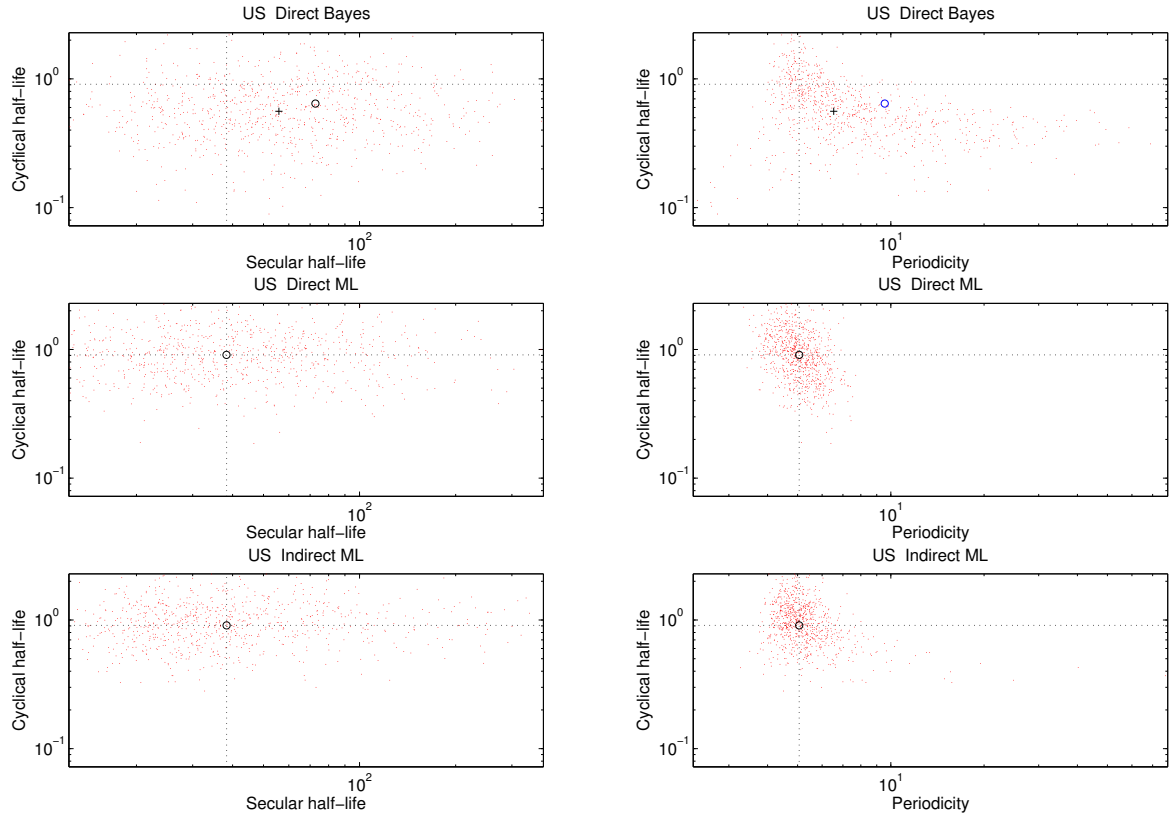

Figure 7. Comparison of posterior and two maximum likelihood asymptotic distributions.

7. Conclusions

The utility of econometric models hinges in no small part on their direct connection to the concepts economists use to create these models and interpret the results of inference. In the example presented in this paper it was possible to replace almost all of the conventional parameters that are convenient for statistical models with such concepts. In many other cases this can be done to partial but significant degree: dynamic stochastic general equilibrium models as utilized in central banks constitute an important class of such examples in macroeconomics. Similar examples in microeconomics and the business sector are also common. If the econometrician is to interpret and communicate findings to the decision-making client, this is important. For the Bayesian econometrician attempting to elicit and incorporate client prior distributions it is even more compelling.

As a practical matter it is equally important for econometricians to be able to bring models to data quickly, avoiding the need for case-by-case special treatments in deriving auxiliary analytical results and tuning computational algorithms. These steps require significant additional time and specialized skills that can place econometric approaches at an overwhelming disadvantage in competition with alternatives (like machine learning) that on their own may not serve the client as well.

The SABL algorithm illustrated in this paper is a tool that addresses these objectives. For a new application, and even for a new model, it presents much lower barriers to entry than do most existing procedures in Bayesian and non-Bayesian econometrics. This is a strong claim, and this paper is a small illustration in support of that claim. Further such illustrations are forthcoming.

Acknowledgments: The Australian Research Council provided financial support through grant DP130103356 and through the ARC Centre of Excellence for Mathematical and Statistical Frontiers of Big Data, Big Models, New Insights, grant CD140100049.

Conflicts of Interest: The author declares no conflict of interest.

References

1. SABL 2015a Handbook, 2015. Available online: http://www.uts.edu.au/sites/default/files/article/downloads/SABL_handbook_2015a.pdf (accessed on 20 February 2016).
2. Geweke, J.; Frischknecht, B. Exact Optimization by Means of Sequentially Adaptive Bayesian Learning. Available online: http://www.uts.edu.au/sites/default/files/article/downloads/WP3.pdf (accessed on 20 February 2016).
3. Durham, G.; Geweke, J. Adaptive sequential posterior simulators for massively parallel computing environments. In *Bayesian Model Comparison (Advances in Econometrics, Volume 34)*; Jeliazkov, I., Poirier, D.J., Eds.; Emerald Group Publishing Limited: West Yorkshire, UK, 2014; Chapter 1; pp. 1–44.
4. Durham, G.; Geweke, J. Sequentially Adaptive Bayesian Learning Algorithms for Inference and Optimization. Available online: http://www.uts.edu.au/sites/default/files/article/downloads/WP5.pdf (accessed on 20 February 2016).
5. Douc, R.; Moulines, E. Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.* **2008**, *36*, 2344–2376.
6. Samuelson, P.A. Interactions between the multiplier analysis and the principle of acceleration. *Rev. Econ. Stat.* **1939**, *21*, 75–78.
7. Geweke, J. The secular and cyclical behavior of real GDP in 19 OECD countries, 1957–1983. *J. Bus. Econ. Stat.* **1988**, *6*, 479–486.
8. Chopin, N. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* **2004**, *32*, 2385–2411.
9. OECD gross domestic product (GDP). Available online: https://stats.oecd.org/index.aspx?queryid=60702 (accessed on 20 February 2016).
10. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by simulated annealing. *Science* **1983**, *220*, 671–680.
11. Goffe, W.L.; Ferrier, G.D.; Rogers, J. Global optimization of statistical functions with simulated annealing. *J. Econom.* **1994**, *60*, 65–99.
12. Zhou, E.; Chen, X. Sequential Monte Carlo simulated annealing. *J. Glob. Optim.* **2013**, *55*, 101–124.