Lecture 2: Simulation methods for state space models

NEIL SHEPHARD
Nuffield College, Oxford, UK
www.nuff.ox.ac.uk/users/shephard/

September 2001
Computationally intensive statistics:
inference for some stochastic processes

1 Statistical models

Many models are just special cases of non-Gaussian, non-linear state space models.

Let us write the state α_t and observations as y_t. States are Markov and the observations are conditionally independent given current state.

Model specified through

$$f(y_t|\alpha_t) \quad \text{and} \quad f(\alpha_{t+1}|\alpha_t).$$

Unfortunately this structure is hard to handle outside Gaussian, linear structure.

- Numerical integration rules — Kitagawa (1987) (high dimensions)

- MCMC — Carlin, Polson, and Stoffer (1992), Carter and Kohn (1994), Fruhwirth-Schnatter (1994), Shephard (1994), Shephard and Pitt (1997) etc (filtering)

- Particle filters — Gordon, Salmond, and Smith (1993), Pitt and Shephard (1999b), Doucet, de Freitas, and Gordon (2001) (likelihood).

2 Classes of state space models: MCMC design

- Unstructured: Markov random field structure only. Carlin, Polson, and Stoffer (1992)

- Conditional Gaussian: $y|s$ is a Gaussian SSF. Carter and Kohn (1994) (s is Markov and discrete), Shephard (1994) (s is Markov).

- Non-Gaussian measurement SSF: ie. $\alpha_{t+1}|\alpha_t$ Gaussian but $f(y_t|Z_t\alpha_t)$ non-Gaussian. Shephard and Pitt (1997).

3 Unstructured SSF

Specify a model through

$$f(y_t|\alpha_t) \quad \text{and} \quad f(\alpha_{t+1}|\alpha_t).$$

so is a special case of a Markov random field.

Inference by simulating from

$$\psi, \alpha|y,$$

where ψ notation for parameters. Carried out in blocks

1. Initialise ψ, α

2. Update
$$\alpha|\psi, y$$

3. Update
$$\psi|\alpha, y.$$

4. Goto 2.

Problem here is $\alpha|\psi, y$ is high dimensional. What to do?

Now

$$\begin{aligned}
& f(\alpha_t|\alpha_1, ..., \alpha_{t-1}, \alpha_{t+1}, ..., \alpha_T, y) \\
= \; & f(\alpha_t|y_t, \alpha_{t+1}, \alpha_{t-1}) \\
\propto \; & f(y_t|\alpha_t) f(\alpha_{t+1}|\alpha_t) f(\alpha_t|\alpha_{t-1}).
\end{aligned}$$

This is $O(1)$ to compute and so MCMC is feasible here. In fact just special case of many imaging problems (MFRs). Carlin, Polson, and Stoffer (1992).

In some cases we can sample from $f(\alpha_t|y_t, \alpha_{t+1}, \alpha_{t-1})$ exactly, e.g. by rejection, or in cases of discrete mixtures.

$$f(\alpha_t|y_t,\alpha_{t+1},\alpha_{t-1}) \propto f(y_t|\alpha_t)f(\alpha_{t+1}|\alpha_t)f(\alpha_t|\alpha_{t-1}).$$

Generic solution is just to make a single site update from some proposal $g(\alpha_t)$ which can depend upon other states and observations. Then use HM to move from α_t^o to

$$\alpha_t^n \sim g(\alpha_t)$$

with probability

$$\begin{aligned} & \min\left\{1, \frac{f(\alpha_t^n|y_t,\alpha_{t+1},\alpha_{t-1})}{f(\alpha_t^o|y_t,\alpha_{t+1},\alpha_{t-1})}\frac{g(\alpha_t^o)}{g(\alpha_t^n)}\right\} \\ = & \min\left\{1, \frac{f(y_t|\alpha_t^n)f(\alpha_{t+1}|\alpha_t^n)f(\alpha_t^n|\alpha_{t-1})}{f(y_t|\alpha_t^o)f(\alpha_{t+1}|\alpha_t^o)f(\alpha_t^o|\alpha_{t-1})}\frac{g(\alpha_t^o)}{g(\alpha_t^n)}\right\}. \end{aligned}$$

Hence all SSF time series can be dealt with this way. Can now use all MCMC tricks on this problem, e.g. random walk proposal, etc.

3.0.1 Gibbs sampler

In some problems it is possible to simulate from $f(\alpha_t|y_t, \alpha_{t+1}, \alpha_{t-1})$, producing the Gibbs sampler. It is rare in time series problems that we can sample from this density exactly. In such cases it is often possible to avoid the above single site updating scheme.

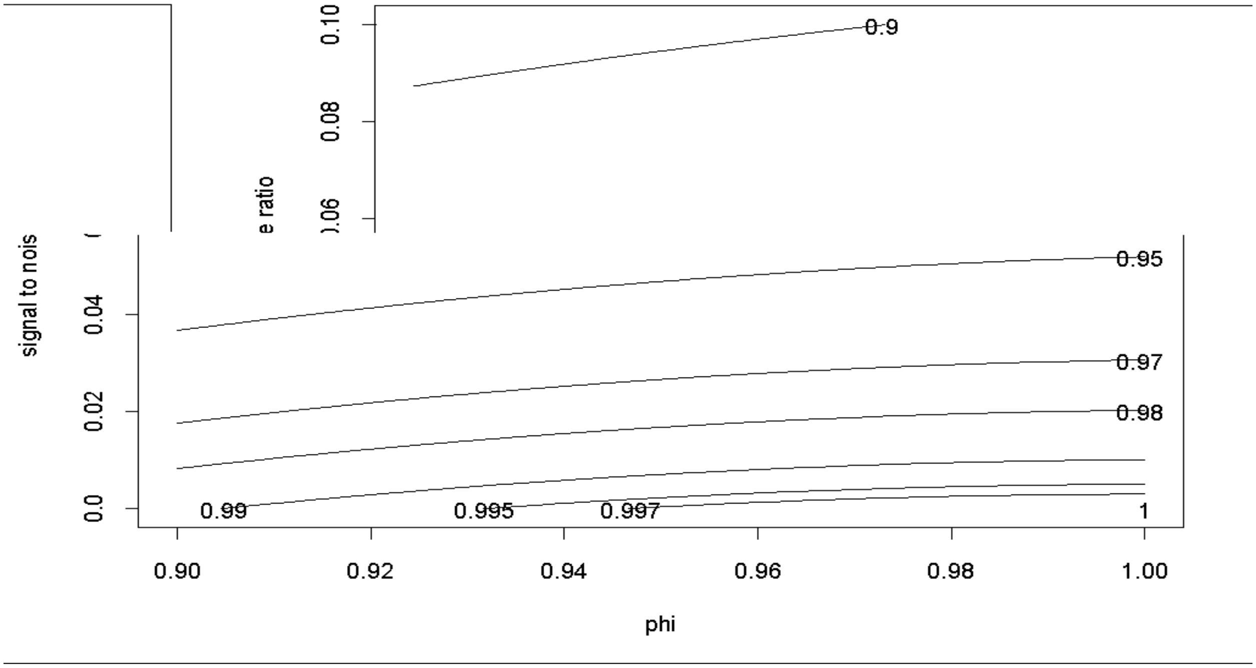
In Pitt and Shephard (1999a) we analytically studied the rate of convergence, using the methods of Roberts and Sahu (1997) in a Gaussian example:

$$\begin{aligned} y_t &= \alpha_t + \varepsilon_t, \qquad & \varepsilon_t &\sim NID(0, \sigma_\varepsilon^2), \\ \alpha_{t+1} &= \phi\alpha_t + \eta_t, \qquad & \eta_t &\sim NID(0, \sigma_\eta^2). \end{aligned}$$

They found, for large n, the (spectral radius) convergence rate was

$$\rho = 4\frac{\phi^2}{\left(1 + \phi^2 + \sigma_\eta^2/\sigma_\varepsilon^2\right)^2},$$

so depends upon signal/nise ratio. As ϕ increases to ρ increases, while as $\sigma_\eta^2/\sigma_\varepsilon^2$ falls so ρ rises. Notice when $\phi = 1$ the expression is particularly simple. In such cases $\sigma_\eta^2/\sigma_\varepsilon^2$ is typically small, which would imply ρ is close to one.



Graph shows a contour plot of convergence rate for various values of the signal to noise ratio and the persistence parameter ϕ.

3.0.2 Parameterisations

Gareth has talked about parameterisation issues. They are important here. Consider putting a mean in the above setup, then

$$\begin{aligned} y_t &= \mu + \alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathsf{NID}(0, \sigma_\varepsilon^2), \qquad t = 1, ..., n, \\ \alpha_t &= \phi\alpha_{t-1} + \eta_t, \qquad \eta_t \sim \mathsf{NID}(0, \sigma_\eta^2), \qquad \alpha_1 \sim \mathsf{N}\left\{0, \sigma_\eta^2/(1-\phi^2)\right\}, \end{aligned}$$

or alternatively

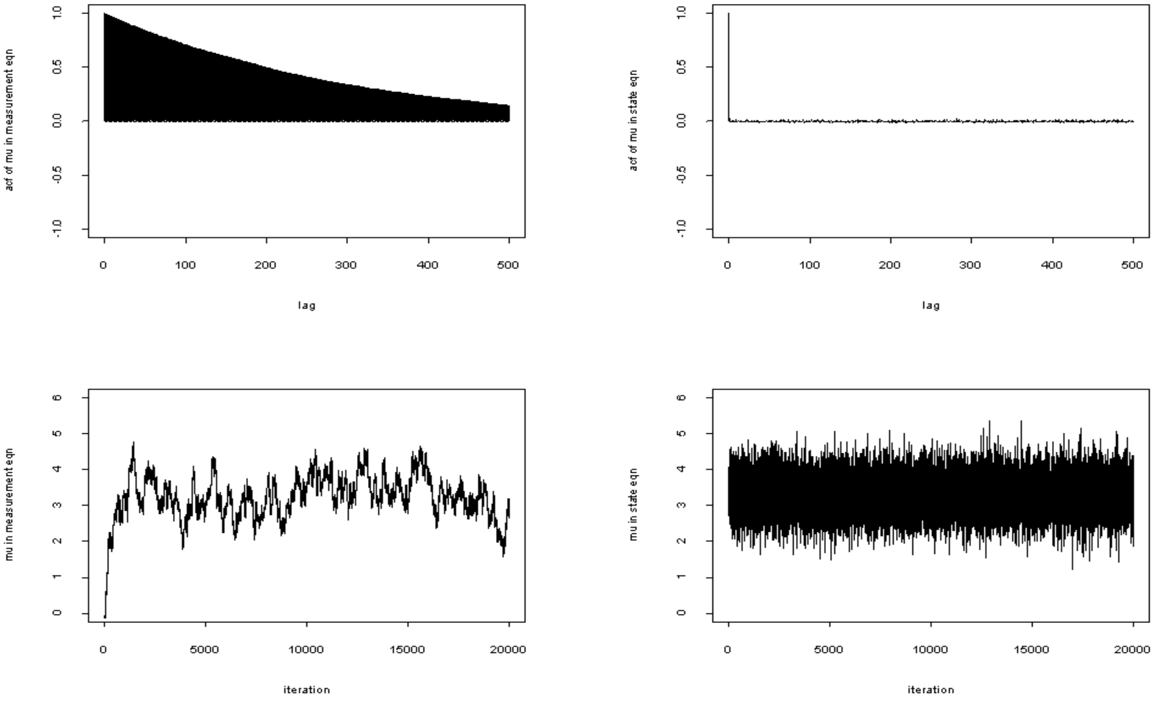
$$\begin{aligned} y_t &= \omega_t + \varepsilon_t,\ \omega_t = \mu + \phi(\omega_{t-1} - \mu) + \eta_t, \text{ and} \\ \omega_1 &\sim \mathsf{N}\left\{\mu, \sigma_\eta^2/(1-\phi^2)\right\}, \quad t = 1, ..., n. \end{aligned}$$

what is the impact?

AR(1) plus noise model above was simulated for $n = 100$ with $\phi = 0.98$, $\sigma_\eta^2 = 0.02$, $\mu = 3.0$ and $\sigma_\varepsilon^2 = 0.1$. Two Gibbs samplers were set up for both the uncentered and centered samplers.

1. $\alpha | y, \mu$ then $\mu | \alpha, y$
2. or $\omega | y, \mu$ then $\mu | \omega$

Performed by using the simulation smoother of de Jong and Shephard (1995). The true relative efficiency is 494.28, whilst the upper bound is 505.05.



Bottom plots are the corresponding sample paths of μ (true value 3), ϕ and σ^2 fixed (at 0.98 and 0.02). The left plots show the case of μ in the measurement equation. The right plots show the corresponding plots for μ in the state equation.

3.0.3 Convenient proposals: the prediction prior

Suppose $\alpha_t^n \sim g(\alpha_t)$ is $f(\alpha_t|\alpha_{t-1})$ then MH becomes

$$\min\left\{1, \frac{f(y_t|\alpha_t^n)f(\alpha_{t+1}|\alpha_t^n)}{f(y_t|\alpha_t^o)f(\alpha_{t+1}|\alpha_t^o)}\right\},$$

which is much simpler. This is typically feasible for most models. Can be a pain to code up the densities.

3.0.4 Convenient proposals: the jackknife prior

Suppose $\alpha_t^n \sim g(\alpha_t)$ is $f(\alpha_t|\alpha_{t-1}, \alpha_{t+1})$ then MH becomes

$$\min\left\{1, \frac{f(y_t|\alpha_t^n)}{f(y_t|\alpha_t^o)}\right\},$$

which is even simpler. This is feasible in models with tractable transition equations, e.g. Gaussian ones or Gaussian mixtures. Many models have this structure. Then only task is to code up measurement equation.

Attractive if state is high dimensional and measurements are of low dimensional.

4 Conditional Gaussian: $y|s$

4.1 Model structure

Perhaps the most "interesting" time series work on this topic is on conditionally Gaussian models. Where

$$y|s$$

is a Gaussian state space model

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t|s \sim NID(0, \Sigma_t), \\ \alpha_{t+1} &= T_t\alpha_t + \eta_t, \quad \eta_t|s \sim NID(0, \Omega_t), \end{aligned}$$

where Z_t, T_t, Σ_t or Ω_t could depend upon s.

Classic models are where s_t is a discrete state Markov chain (Carter and Kohn (1994)) or more general Markov processes (Shephard (1994)).

4.2 Outlier/skew models

e.g.

$$\begin{array}{rcll} y_t & = & \alpha_t + \varepsilon_t, & \varepsilon_t | s_t \sim NID(\beta s_{1t}, s_{1t}\sigma_\varepsilon^2), \\ \alpha_{t+1} & = & \phi \alpha_t + \eta_t, & \eta_t | s_t \sim NID(\gamma s_{2t}, s_{2t}\sigma_\eta^2). \end{array}$$

Then we allow $s_{1t} \perp\!\!\!\perp s_{2t}$. Example of this where there is no dependence amoungst the s_t :

- $\beta = \gamma = s_{2t} = 0$ and s_{1t} has 2 values. One is large, the other small, then unconditionally ε_t is a mixture of 2 normals. Large tradition in time series.

- $\gamma = s_{2t} = 0$ and s_{1t} has 2 values. Above but allows skewness. Suppose $\beta < 0$, then can allow ε_t to have a negative skew.

- $\beta = \gamma = 0$ and s_{1t}, s_{2t} each has 2 values allows both innovative and transitory outliers.

- s_{1t}, s_{2t} are independent generalised inverse Gaussian then $\varepsilon_t \perp\!\!\!\perp \eta_t$ are iid generalised hyperbolic. Examples include student t, Laplace, normal gamma and normal inverse Gaussian distributions. Model based outliers.

Traditionally this type of model has been handled using ad hoc methods. Above give coherent robust signal extraction. e.g. robust seasonal adjustment.

Attractive as within the recent Bayesian tradition of nonparametric density estimation via mixtures e.g. Richardson and Green (1997).

4.3 Discrete state Markov chains

e.g.

$$\begin{array}{lcll} y_t & = & \beta s_t + \alpha_t + \varepsilon_t, & \varepsilon_t \sim NID(0, \sigma_\varepsilon^2), \\ \alpha_{t+1} & = & \phi \alpha_t + \eta_t, & \eta_t \sim NID(0, \sigma_\eta^2). \end{array}$$

allows the level in the process to follow discrete state Markov chain. In economics we might think of s_t as indicating recession and upswing.

4.4 Multiplicative models

e.g.

$$\begin{aligned}
y_t &= s_t\alpha_t + \varepsilon_t, & \varepsilon_t &\sim NID(0, \sigma^2_\varepsilon), \\
\alpha_{t+1} &= \phi_1\alpha_t + \eta_t, & \eta_t &\sim NID(0, \sigma^2_\eta), \\
s_{t+1} &= \phi_1\alpha_t + e_t & e_t &\sim NID(0, \sigma^2_e),
\end{aligned}$$

4.5 Computations

Attractive block structure of these conditionally Gaussian models. Sample from

$$\alpha, s|y$$

by

1. Initialise s

2. Update

   $$\alpha|y, s,$$

   is a multivariate Gaussian density. We will see we can sample from this quickly in a moment.

3. Update

   $$s|y, \alpha.$$

   If s_t is discrete/Markov then we can, in general, draw from $s|y, \alpha$ in one block (Carter and Kohn (1994)). Otherwise we have to sample from

   $$s_t|s_{t-1}, s_{t+1}, \alpha_t, \alpha_{t-1}, \alpha_{t+1}, y_t$$

   perhaps using a MH update.

4.6 Simulation smoother

A major, time series, advance occured in 1994 with the publication of simulation smoothers. These are generic algorithms for simulating from the posteriod distribution of the states given the data. In particular $\alpha|y$ where

$$\begin{aligned} y_t &= Z_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \Sigma_t), \\ \alpha_{t+1} &= T_t\alpha_t + \eta_t, \quad \eta_t \sim NID(0, \Omega_t), \end{aligned}$$

Initial results due to Carter and Kohn (1994) and Fruhwirth-Schnatter (1994). Later results which gave more computationally efficient algorithms were de Jong and Shephard (1995) and recently Durbin and Koopman (2002). The algorithmic details are rather dull from the perspective of this course, just some linear algebra.

Software to carry this out is available in SsfPack (based in Ox) and is currently being ported to Splus.

5 Non-Gaussian measurement TS

Maintain the dynamic structure

$$\alpha_{t+1} = T_t\alpha_t + \eta_t, \quad \eta_t \sim NID(0, \Omega_t),$$

but now allow a link function $\theta_t = g(Z_t\alpha_t)$ where

$$f(y_t|\theta_t).$$

Important examples of this setup include allowing binary, count or non-negative time series.

e.g. generalised linear time series models

$$l(\theta_t) = \log f(y_t|\theta_t) = y_t\theta_t - b(\theta_t) + c(y_t).$$

Can make Laplace type proposals based upon blocks. Studied in Shephard and Pitt (1997).

Approximate $l(\theta_t)$ by

$$l(\widehat{\theta}_t) + z_t(\alpha_t - \widehat{\alpha}_t)l'(\widehat{\theta}_t) + \frac{1}{2}\left\{z_t(\alpha_t - \widehat{\alpha}_t)\right\}^2 D_t(\widehat{\theta}_t).$$

That is a Gaussian approximation. Then add to Gaussian transition. Can do this for a single observation at a time or over a block. Many models take

$$D_t(\widehat{\theta}_t) = -\ddot{b}\,(\widehat{\theta}_t) < 0.$$

Allows easy proposal for MH. Notice the MH acceptance rate only depends upon the measurement density.

6 Particle filters

6.1 Basics

MCMC does not do filtering — only smoothing & parameter estimation. What to do?

Need filtering because

- substantial interest from a subject matter perspective
- construct diagnostics, e.g. $F(y_{t+1}|\mathcal{F}_t)$. They are standard uniform, iid, if model is true.

Filtering recursion is

$$f(\alpha_{t+1}|\mathcal{F}_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) \int f(\alpha_{t+1}|\alpha_t) dF(\alpha_t|\mathcal{F}_t).$$

Replace

$$f(\alpha_{t+1}|\mathcal{F}_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) \int f(\alpha_{t+1}|\alpha_t) dF(\alpha_t|\mathcal{F}_t).$$

by empirical version

$$\widehat{f}(\alpha_{t+1}|\mathcal{F}_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) \sum_{j=1}^{M} f(\alpha_{t+1}|\alpha_t^j),$$

where α_t^j is a sample from $F(\alpha_t|\mathcal{F}_t)$. Now as $M \to \infty$ good approximation.

Task: sample from

$$\widehat{f}(\alpha_{t+1}|\mathcal{F}_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) \sum_{j=1}^{M} f(\alpha_{t+1}|\alpha_t^j).$$

Lot of work on this: Gordon, Salmond, and Smith (1993), Kitagawa (1996), Berzuini, Best, Gilks, and Larizza (1997) and Isard and Blake (1996). Doucet, de Freitas, and Gordon (2001).

Sample from

$$\widehat{f}(\alpha_{t+1}|\mathcal{F}_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) \sum_{j=1}^{M} f(\alpha_{t+1}|\alpha_t^j),$$

by sampling from

$$\widehat{f}(\alpha_{t+1}, j|\mathcal{F}_{t+1}) \propto f(y_{t+1}|\alpha_{t+1}) f(\alpha_{t+1}|\alpha_t^j),$$

then throw away j. Yields required samples. Can use SIR or MCMC now. Often can sample from this directly.

References

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

Carlin, B. P., N. G. Polson, and D. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modelling. *Journal of the American Statistical Association 87*, 493–500.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*, 541–53.

de Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*, 339–50.

Doucet, A., N. de Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis 15*, 183–202.

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

Isard, M. and A. Blake (1996). Contour tracking by stochastic propagation of conditional density. *Proceedings of the European Conference on Computer Vision, Cambridge 1*, 343–356.

Kitagawa, G. (1987). Non-Gaussian state space modelling of non-stationary time series. *Journal of the American Statistical Association 82*, 503–514.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Computational and Graphical Statistics 5*, 1–25.

Pitt, M. K. and N. Shephard (1999a). Analytic convergence rates and parameterisation issues for the Gibbs sampler applied to state space models. *Journal of Time Series Analysis 21*, 63–85.

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

Richardson, S. and P. Green (1997). On Bayesian analysis of mixtures with unknown number of components. *Journal of the Royal Statistical Society, Series B 59*, 731–92.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B 59*, 291–317.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–31.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–67.