Inference for dynamic and latent variable models via iterated, perturbed Bayes maps

Edward Ionides

Department of Statistics, University of Michigan

Statistics Colloquium
Harvard University

Tuesday February 1, 2016

Slides for this talk are at
`http://dept.stat.lsa.umich.edu/~ionides/talks/harvard16.pdf`

Overview

1. Introduction to iterated filtering methodology.
2. A new iterated filtering algorithm (IF2).
3. Theoretical justification of IF2.
4. Applications of IF2.

Partially observed Markov process (POMP) models

- Data $y_1^*, \dots, y_N^*$ collected at times $t_1 < \dots < t_N$ are modeled as noisy and incomplete observations of a Markov process $\{X(t), t \geq t_0\}$.
- This is a **partially observed Markov process (POMP)** model, also known as a hidden Markov model or a state space model.
- The POMP model may depend on an unknown parameter vector, θ.
- Scientific uses of POMP models are too numerous to list. They include many applications to biological systems, rocket science, economics, geophysical systems, etc.

Sequential Monte Carlo (SMC) methods for POMP models

- **Filtering** is estimation of the latent dynamic process $X(t_n)$ given data $y_1^*, \dots, y_N^*$ for a fixed POMP model, i.e., with parameter θ assumed known.
- **Sequential Monte Carlo (SMC)** is a numerical method for filtering and evaluating the likelihood function.
- SMC is also called a "particle filter."
- Filtering has extensive applications in science and engineering. Over the last 15 years, SMC has become popular for filtering non-linear, non-Gaussian partially observed Markov process (POMP) models.

A brief history of Monte Carlo methods

- The basic Monte Carlo method approximates $\int h(x)\,f(x)\,dx \approx \frac{1}{J}\sum_{j=1}^{J} h(X_j), \quad \text{where } X_j \sim f.$
- For high-dimensional integration, Markov chain Monte Carlo (MCMC) draws $X_j \sim f$ by setting up a Markov chain with stationary distribution f.
- Sequential Monte Carlo (SMC) breaks down the high-dimensional integration problem into a sequence of lower-dimension problems.
- SMC was called "Poor man's Monte Carlo" by Hammersley (1954). The theory and practice of SMC is now comparable to MCMC.
- To simulate the 3D structure of a molecule with 1000 atoms, MCMC transitions adjust the position of all 1000 atoms; SMC builds up the molecule one atom at a time.

What is iterated filtering?

- Iterated filtering algorithms adapt sequential Monte Carlo (SMC) into a tool for inference on the unknown parameter vector, θ.
- We call IF1 the iterated filtering algorithm of Ionides, Bretó & King (2006). IF1 uses an extended POMP model where θ is replaced by a time-varying process $\theta(t)$ which follows a random walk. SMC filtering on this model can approximate the derivative of the log likelihood.
- Novel algorithms were needed because two “obvious” approaches to parameter inference via SMC fail in all but simple problems:
  - Apply a black-box optimizer such as Nelder-Mead to the SMC evaluation of the likelihood.
  - Carry out Bayesian inference by SMC with θ added to the POMP as a static parameter.
- Standard MCMC and EM algorithms also work poorly on POMP models.

Dobson, A. (2014). Mathematical models for emerging disease. *Science* 346:1294–1295

“Powerful new inferential fitting methods (Ionides, Bretó and King, 2006) considerably increase the accuracy of outbreak predictions while also allowing models whose structure reflects different underlying assumptions to be compared. These approaches move well beyond time series and statistical regression analyses as they include mechanistic details as mathematical functions that define rates of loss of immunity and the response of vector abundance to climate.”

Notation for partially observed Markov process models

- Write $X_n = X(t_n)$ and $X_{0:N} = (X_0, \ldots, X_N)$. Let Y_n be a random variable modeling the observation at time t_n.
- The one-step transition density, $f_{X_n|X_{n-1}}(x_n \,|\, x_{n-1}\,;\theta)$, together with the measurement density, $f_{Y_n|X_n}(y_n \,|\, x_n\,;\theta)$ and the initial density, $f_{X_0}(x_0\,;\theta)$, specify the entire joint density via

$$f_{X_{0:N},Y_{1:N}}(x_{0:N}, y_{1:N};\theta) = f_{X_0}(x_0;\theta) \prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)\, f_{Y_n|X_n}(y_n|x_n;\theta).$$

- The likelihood function is

$$\ell(\theta) = f_{Y_{1:N}}(y^*_{1:N}\,;\theta) = \int f_{X_{0:N},Y_{1:N}}(x_{0:N}, y^*_{1:N};\theta)\, dx_{0:N}$$

Input and output for the IF2 algorithm

input:

Simulator for latent process initial density, $f_{X_0}(x_0\,;\theta)$

Simulator for transition density, $f_{X_n|X_{n-1}}(x_n\,|\,x_{n-1}\,;\theta)$, n in $1\!:\!N$

Evaluator for measurement density, $f_{Y_n|X_n}(y_n\,|\,x_n\,;\theta)$, n in $1\!:\!N$

Data, $y^*_{1:N}$

Number of iterations, M

Number of particles, J

Initial parameter swarm, $\{\Theta^0_j,\ j \text{ in } 1\!:\!J\}$

Perturbation density, $h_n(\theta\,|\,\varphi\,;\sigma)$, n in $1\!:\!N$

Perturbation sequence, $\sigma_{1:M}$

output: Final parameter swarm, $\{\Theta^M_j,\ j \text{ in } 1\!:\!J\}$

Algorithms that specify the dynamic model via a simulator are said to be **plug-and-play**. This property ensures applicability to the broad class of models for which a simulator is available.

IF2: iterated SMC with perturbed parameters

```
For $m$ in $1\!:\!M$ [$M$ filtering iterations, with decreasing $\sigma_m$]
    $\Theta^{F,m}_{0,j} \sim h_0(\,\cdot \mid \Theta^{m-1}_j\,; \sigma_m)$ for $j$ in $1\!:\!J$
    $X^{F,m}_{0,j} \sim f_{X_0}(x_0; \Theta^{F,m}_{0,j})$ for $j$ in $1\!:\!J$
    For $n$ in $1\!:\!N$ [SMC with $J$ particles]
        $\Theta^{P,m}_{n,j} \sim h_n(\,\cdot \mid \Theta^{F,m}_{n-1,j}, \sigma_m)$ for $j$ in $1\!:\!J$
        $X^{P,m}_{n,j} \sim f_{X_n|X_{n-1}}(x_n \mid X^{F,m}_{n-1,j}; \Theta^{P,m}_j)$ for $j$ in $1\!:\!J$
        $w^m_{n,j} = f_{Y_n|X_n}(y^*_n \mid X^{P,m}_{n,j}; \Theta^{P,m}_{n,j})$ for $j$ in $1\!:\!J$
        Draw $k_{1:J}$ with $\mathbb{P}(k_j = i) = w^m_{n,i} \Big/ \sum_{u=1}^J w^m_{n,u}$
        $\Theta^{F,m}_{n,j} = \Theta^{P,m}_{n,k_j}$ and $X^{F,m}_{n,j} = X^{P,m}_{n,k_j}$ for $j$ in $1\!:\!J$
    End For
    Set $\Theta^m_j = \Theta^{F,m}_{N,j}$ for $j$ in $1\!:\!J$
End For
```

IF2: iterated SMC with perturbed parameters

```
For $m$ in $1\!:\!M$
    $\Theta^{F,m}_{0,j} \sim h_0(\,\cdot \mid \Theta^{m-1}_j\,; \sigma_m)$ for $j$ in $1\!:\!J$
    $X^{F,m}_{0,j} \sim f_{X_0}(x_0; \Theta^{F,m}_{0,j})$ for $j$ in $1\!:\!J$
    [carry out SMC on an extended model, with the time-varying
    parameters included in the latent state, initialized at $(X^{F,m}_{0,j}, \Theta^{F,m}_{0,j})$]
    Set $\Theta^m_j = \Theta^{F,m}_{N,j}$ for $j$ in $1\!:\!J$
End For
```

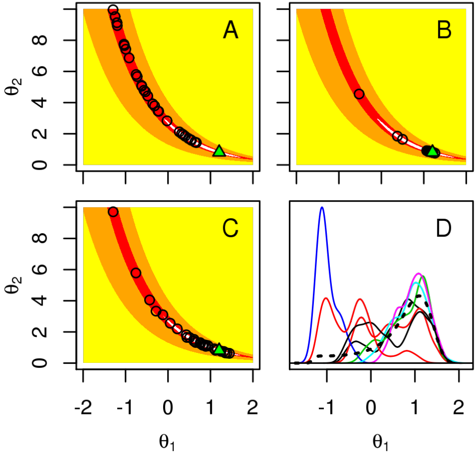
Numerical examples

- We compare IF1, IF2 and the particle Markov chain Monte Carlo (PMCMC) method of Andrieu et al (2010).
- PMCMC is an SMC-based plug-and-play algorithm for full-information Bayesian inference on POMPs.
- Computations were done using the `pomp` R package:

  King, Nguyen & Ionides (2015). "Statistical inference for partially observed Markov processes via the R package `pomp`." To appear in *Journal of Statistical Software*. Available at `http://kingaa.github.io/pomp/vignettes/pompjss.pdf`

- Data and code reproducing our results are a supplement to

  Ionides, Nguyen, Atchadé, Stoev & King (2015). "Inference for dynamic and latent variable models via iterated, perturbed Bayes maps." *Proceedings of the National Academy of Sciences of the USA*.



Toy example.

$X(t) = \big(\exp\{\theta_1\}, \theta_2 \exp\{\theta_1\}\big)$,

constant for all t.

100 independent observations:

Given $X(t) = x$,

$$Y_n \sim \mathrm{Normal}\left[x, \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}\right].$$

A. IF1 point estimates from 30 replications and the MLE (green triangle).
B. IF2 point estimates from 30 replications and the MLE (green triangle).
C. Final parameter value of 30 PMCMC chains with 10^4 filtering iterations.
D. Kernel density estimates from 8 of these 30 PMCMC chains, and the true posterior distribution (dotted black line).

Why is IF2 so much better than IF1 on this problem?

- IF1 updates parameters by a linear combination of filtered parameter estimates for the extended model with time-varying parameters.
- Taking linear combinations can knock the optimizer off nonlinear ridges of the likelihood function.
- IF2 does not have this vulnerability.
- A heuristic argument suggests that IF2 has 2nd order convergence (as if the 1st and 2nd derivatives were computable) whereas IF1 has 1st order convergence. It is an open problem to formalize that.

Epidemiological applications: A review of disease dynamics

- Communicable diseases have long had major global health impact (malaria, tuberculosis, measles, etc).
- Emerging diseases need to be understood and controlled (HIV, Ebola, bird flu, SARS, etc).
- Central to math models is an infected population, $I(t)$, which interacts with a susceptible population, $S(t)$. Susceptible individuals become infected at a nonlinear rate $\beta\, I(t)\, S(t)$, where β is a contact rate.
- The inherent stochasticity of biological populations, and our partial ability to observe epidemics, therefore lead to nonlinear POMP model inference problems.

Application to a cholera model

The study population $P(t)$ is split into susceptibles, $S(t)$, infecteds, $I(t)$, and k recovered classes $R_1(t), \ldots, R_k(t)$. The state process $X(t) = (S(t), I(t), R_1(t), \ldots, R_k(t))$ follows a stochastic differential equation driven by a Brownian motion $\{B(t)\}$,

$$\begin{aligned}
dS &= \big\{k\epsilon R_k + \delta(P - S) - \lambda(t)\, S\big\}dt \,+ dP - (\sigma I/P)\, dB, \\
dI &= \big\{\lambda(t)\, S - (m + \delta + \gamma)I\big\}dt + (\sigma I/P)\, dB, \\
dR_1 &= \big\{\gamma I - (k\epsilon + \delta)R_1\big\}dt, \\
&\vdots \\
dR_k &= \big\{k\epsilon R_{k-1} - (k\epsilon + \delta)R_k\big\}dt.
\end{aligned}$$

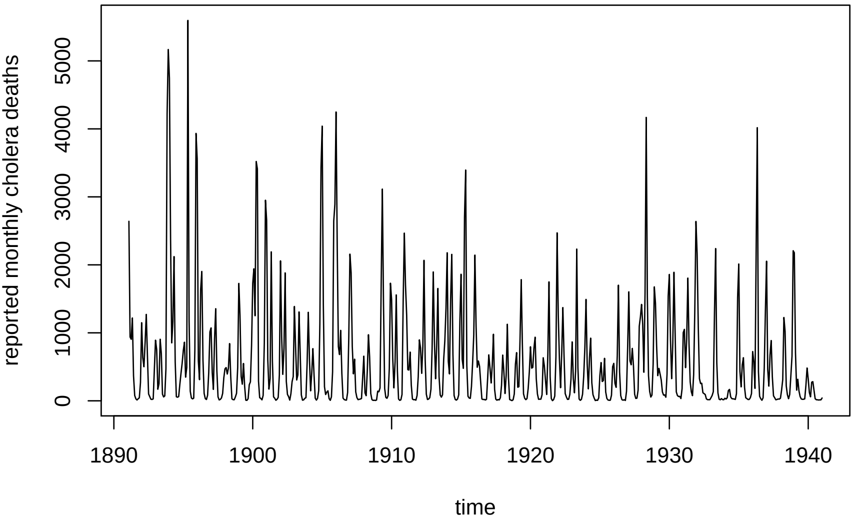
The nonlinearity arises through the force of infection, $\lambda(t)$, specified as

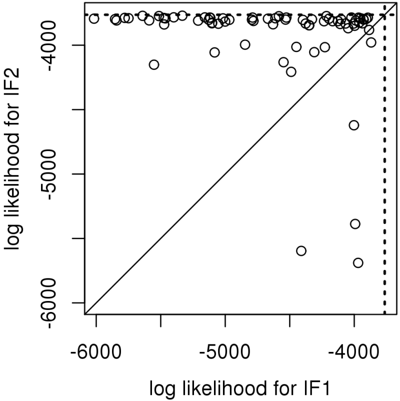
$$\lambda(t) = \bar{\beta} \exp\Big\{\beta_{\mathrm{trend}}(t - t_0) + \textstyle\sum_{j=1}^{N_s} \beta_j s_j(t)\Big\} I/P + \bar{\omega} \exp\Big\{\textstyle\sum_{j=1}^{N_s} \omega_j s_j(t)\Big\},$$

where $\{s_j(t), j = 1, \ldots, N_s\}$ is a periodic cubic B-spline basis. The data are monthly counts of cholera mortality, modeled as

$$Y_n \sim \mathrm{Normal}(M_n, \tau^2 M_n^2) \text{ for } M_n = \textstyle\int_{t_{n-1}}^{t_n} m\, I(s)\, ds.$$

Monthly cholera mortality in Dhaka, 1891-1940





Comparison of IF1 and IF2 on the cholera model.

Algorithmic tuning parameters for both IF1 and IF2 were set at the values chosen by King et al (2008) for IF1.

- Log likelihoods of the parameter vector output by IF1 and IF2, both started at a uniform draw from a large 23-dimensional hyper-rectangle.
- Dotted lines show the maximum log likelihood.

IF2 as an iterated Bayes map

- Each iteration of IF2 is a Monte Carlo approximation to a map
$$T_\sigma f(\theta_N) = \frac{\int \breve{\ell}(\theta_{0:N}) h(\theta_{0:N}|\varphi\,;\sigma) f(\varphi)\, d\varphi\, d\theta_{0:N-1}}{\int \breve{\ell}(\theta_{0:N}) h(\theta_{0:N}|\varphi\,;\sigma) f(\varphi)\, d\varphi\, d\theta_{0:N}}, \qquad (1)$$
where $\breve{\ell}(\theta_{0:N})$ is the likelihood of the data under the extended model with time-varying parameter $\theta_{0:N}$.
- f and $T_\sigma f$ in (1) approximate the initial and final density of the IF2 parameter swarm.
- When the standard deviation of the parameter perturbations is held fixed at $\sigma_m = \sigma > 0$, IF2 is a Monte Carlo approximation to $T_\sigma^M f(\theta)$.
- Iterated Bayes maps are not usually contractions.
- We study the homogeneous case, $\sigma_m = \sigma$.
- Studying the limit $\sigma \to 0$ may be as appropriate as an asymptotic analysis to study the practical properties of a procedure such as IF2, with σ_m decreasing down to some positive level $\sigma > 0$ but never completing the asymptotic limit $\sigma_m \to 0$.

IF2 as a generalization of data cloning

- In the case $\sigma = 0$, the iterated Bayes map corresponds to the data cloning approach of Lele (2007).
- For $\sigma = 0$, Lele et al (2007) found central limit theorems. For $\sigma \neq 0$, the limit as $M \to \infty$ is not usually Gaussian.
- Taking $\sigma \neq 0$ adds numerical stability, which is necessary for convergence of SMC approximations.

Theorem 1. Assuming adequate regularity conditions, there is a unique probability density f_σ with

$$\lim_{M\to\infty} T_\sigma^M f = f_\sigma,$$

with the limit taken in the L^1 norm. The SMC approximation to $T_\sigma^M f$ converges to $T_\sigma^M f$ as $J \to \infty$, uniformly in M.

- Theorem 1 follows from existing results on filter stability.
- Convergence and stability of the ideal filter (a small error at time t has diminishing effects at later times) is closely related to convergence of SMC.

Theorem 2. Under regularity conditions, $\lim_{\sigma \to 0} f_\sigma$ approaches a point mass at the maximum likelihood estimate (MLE).

Outline of proof.

- Trajectories in parameter space which stray away from the MLE are down-weighted by the Bayes map relative to trajectories staying close to the MLE.
- As σ decreases, excursions any fixed distance away from the MLE require an increasing number of iterations and therefore receive an increasing penalty from the iterated Bayes map.
- Bounding this penalty proves the theorem.

Conclusions

- IF1 enabled previously unfeasible likelihood-based inference for nonlinear, non-Gaussian POMP models.
- We have not yet found a situation where IF2 performs worse than IF1. In complex nonlinear models, we have found IF2 always substantially better.
- In addition, IF2 is simpler. Some extensions are easier: IF2 can readily handle parameters for which the information in the data is concentrated in a sub-interval.
- If you like IF1, you'll love IF2.
- IF2, together with advances in software and hardware, makes inference for nonlinear POMP models readily accessible to Masters level statisticians. Evidence: a short coure on Simulation-based Inference for Epidemiological Dynamics (`http://kingaa.github.io/sbied`).

Current and future work

- Existing sequential Monte Carlo (SMC) methods fail in high-dimensional systems, such as space-time modeling. Can this be resolved? Does iterated filtering help with inference?
- Genetic sequence data on pathogens should be informative about disease transmission dynamics. This challenge leads to models that are not quite POMPs, but SMC and IF2 are applicable.
- The iterated Bayes map results underlying IF2 apply to general latent variable models, not just POMPs. Could IF2-like algorithms assist challenges such as inference for big hierarchical random effect models?
- Big data: nowadays one often has many time series. Iterated filtering extends to mechanistic models for panel, or longitudinal, data:

  > Romero-Severson et al. (2015). “Dynamic variation in sexual contact rates for a cohort of HIV-negative urban gay men.” *American Journal of Epidemiology*.

  A full theory and methodology for this situation is in development.

Andrieu, C., Doucet, A., and Holenstein, R. (2010).
Particle Markov chain Monte Carlo methods.
J. R. Stat. Soc. B, 72:269–342.

Ionides, E. L., Bretó, C., and King, A. A. (2006).
Inference for nonlinear dynamical systems.
Proc. Natl. Acad. Sci. USA, 103:18438–18443.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015).
Inference for dynamic and latent variable models via iterated, perturbed Bayes maps.
Proc. Natl. Acad. Sci. USA.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008).
Inapparent infections and cholera dynamics.
Nature, 454:877–880.

Lele, S. R., Dennis, B., and Lutscher, F. (2007).
Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods.
Ecology Letters, 10(7):551–563.

More iterated filtering references can be found on Wikipedia

`wikipedia.org/wiki/Iterated_filtering`

Thank You!

The End.