Piecewise Approximate Bayesian Computation

Fast inference for discretely observed Markov models

Theo Kypraios
http://www.maths.nott.ac.uk/∼tk

From Spectral Gaps to Particle Filters, Reading, September, 2013

Joint work:

- Simon Preston (Universty of Nottingham)
- Simon White (Biostatistics Unit, MRC @ Cambridge)

More details in:

`http://arxiv.org/abs/1301.2975`.

Elements of Statistical Inference

- *Different focus* compared to previous talks.
- **Motivation**:
  1. Given some observed data ...
  2. ... that we assume to have been generated by a (stochastic) model ...
  3. ... we wish to make inference for the model parameters.
- Interest is in cases where conventional methods fail, computationally expensive, or simply don't work ....
- Having a good understanding of the model dynamics is of great importance when designing new methods/algorithms.

Elements of Statistical Inference

- *Different focus* compared to previous talks.
- **Motivation**:
  1. Given some observed data . . .
  2. . . . that we assume to have been generated by a (stochastic) model . . .
  3. . . . we wish to make inference for the model parameters.
- Interest is in cases where conventional methods fail, computationally expensive, or simply don't work . . ..
- Having a good understanding of the model dynamics is of great importance when designing new methods/algorithms.

Elements of Statistical Inference

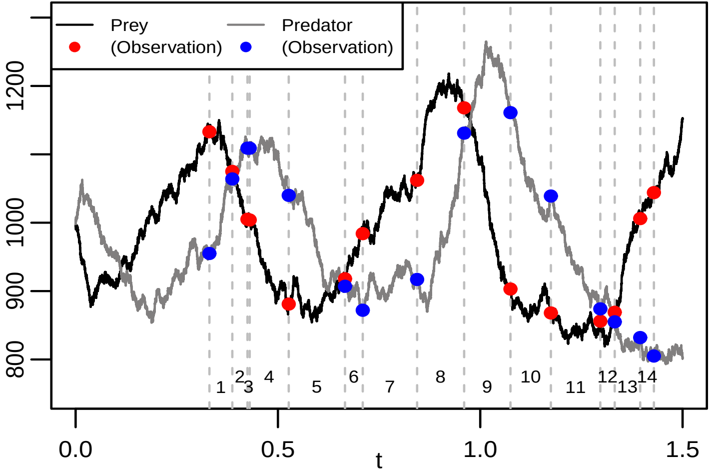
- *Different focus* compared to previous talks.
- **Motivation**:
  1. Given some observed data ...
  2. ... that we assume to have been generated by a (stochastic) model ...
  3. ... we wish to make inference for the model parameters.
- Interest is in cases where conventional methods fail, computationally expensive, or simply don't work ....
- Having a good understanding of the model dynamics is of great importance when designing new methods/algorithms.

Elements of Statistical Inference

- *Different focus* compared to previous talks.
- **Motivation**:
  1. Given some observed data ...
  2. ... that we assume to have been generated by a (stochastic) model ...
  3. ... we wish to make inference for the model parameters.
- Interest is in cases where conventional methods fail, computationally expensive, or simply don't work ....
- Having a good understanding of the model dynamics is of great importance when designing new methods/algorithms.

A Motivating Example: Lotka–Volterra

Suppose our data are a set of observations denoted $\mathcal{X} = \{x_1, \ldots, x_n\} = \{x(t_1), \ldots, x(t_n)\}$ of state variable $x \in \mathbb{R}^m$ at time points $t_1, \ldots, t_n$.



Elements of Statistical Inference (2)

- Denote by $\mathcal{X}$ our observed data and by θ the parameter(s) of interest.
- If we were to employ a *maximum likelihood* approach then we must be able to write down/evaluate the likelihood $\pi(\mathcal{X}|\theta)$ i.e. the probability of observing the data $\mathcal{X}$ what we have observed) for all parameter values θ ...
- ... and then find which parameter(s) θ maximise the likelihood.
- Use asymptotic theory and obtain (approximate) confidence intervals for θ to quantify uncertainty.
- In this talk we adopt a Bayesian framework.

Exact Bayesian Computation (EBC)

- Suppose we have discrete data $\mathcal{X}$, prior $\pi(\theta)$ for parameter(s) θ.
- Aim: Draw samples from the posterior distribution of the parameters, $\pi(\theta|\mathcal{X})$.
- Bayes Theorem gives

$$\pi(\theta|\mathcal{X}) = \pi(\mathcal{X}|\theta)\pi(\theta)/\pi(\mathcal{X})$$

where $\pi(\mathcal{X})$ is a normalising constant:

$$\pi(\mathcal{X}) = \int_\theta \pi(\mathcal{X}|\theta)\pi(\theta)\ \mathrm{d}\theta$$

and therefore

$$\pi(\theta|\mathcal{X}) \propto \pi(\mathcal{X}|\theta)\pi(\theta)$$

Rejection Sampling

Consider the following algorithm:

Algorithm 1
Exact Bayesian Computation (EBC)

1: Sample θ^* from $\pi(\theta)$.
2: Simulate dataset x^* from the model using parameters θ^*.
3: Accept θ with probability equal to $\pi(\mathcal{X}|\theta)$
4: Repeat.

---

* Note that this algorithm requires that we are able to compute the likelihood, $\pi(D|\theta)$, any θ (Step 2).

Rejection Sampling

Consider the following algorithm:

Algorithm 2

Exact Bayesian Computation (EBC)

1: Sample θ^* from $\pi(\theta)$.

2: Simulate dataset x^* from the model using parameters θ^*.

3: Accept θ^* if $x^* = x$, otherwise reject.

4: Repeat.

---

* Algorithm 2 is equivalent to Algorithm 1.

* Evaluating $\pi(\mathcal{X}|\theta) \iff$ *simulate* an event which occurs with that probability.

* That means that the calculation of the likelihood is unnecessary as long as we can simulate from our stochastic model.

Approximate Bayesian Computation

Algorithm 2 is only of practical use if $\mathcal{X}$ is discrete, else the acceptance probability in Step 3 is zero.

For continuous distributions Pritchard *et al.* (1999) suggested the following algorithm:

Algorithm 3
Approximate Bayesian Computation (ABC)

As Algorithm 2, but with step 3 replaced by:

3': Accept θ^* if $d\big(s(\mathcal{X}), s(\mathcal{X}^*)\big) \leq \varepsilon$, otherwise reject.

---

where $d(\cdot,\cdot)$ is a distance function, usually taken to be the L^2-norm of the difference between its arguments; $s(\cdot)$ is a function of the data; and ε is a tolerance.

Approximate Bayesian Computation

- In practice iss rarely possible to use an $s(\cdot)$ which is sufficient, or to take ε especially small (or zero).
- ABC requires a careful choice of $s(\cdot)$ and ε to make the acceptance rate tolerably large, at the same time as trying not to make the ABC posterior too different from the true posterior, $\pi(\theta|\mathcal{X})$.
- Over the last decade, a wide range of extensions to the original ABC algorithm have been developed (MCMC-ABC, SMC-ABC, Semi-Automatic ABC ...)
- ..., however, computational cost remains a central issue since it determines the balance that can be made between Monte Carlo error/bias (via summary stats).

Piecewise Approximate Bayesian Computation (PW-ABC)

- Interested in exploring cases (i.e. models/data) and methods where *ideally*, exact Monte-Carlo inference can be drawn in practice without having to compute likelihoods either because it is
  - too expensive to compute
  - or, intractable;
- or, difficult to maximise or sample from the posterior distribution of interest.
- If *exact* inference seems infeasible $\rightarrow$ efficient, but approximate likelihood-free inference.
- A guiding principle is to take every opportunity to exploit model structure to minimize computational costs.

Exploting the Markovian Structure

The Markov property enables the likelihood to be written as

$$\begin{aligned}\pi(\mathcal{X}|\theta) &= \left(\prod_{i=2}^{n} \pi(x_i|x_{i-1},\ldots,x_1,\theta)\right)\pi(x_1|\theta)\\ &= \left(\prod_{i=2}^{n} \pi(x_i|x_{i-1},\theta)\right)\pi(x_1|\theta), \qquad (1)\end{aligned}$$

if n is large then we can ignore the contribution of the first data point (x_1) to the likelihood, and write

$$\pi(\mathcal{X}|\theta) = \prod_{i=2}^{n} \pi(x_i|x_{i-1},\theta)$$

Hence the posterior as

$$
\begin{aligned}
\pi(\theta|\mathcal{X}) &\propto \pi(\theta)\cdot\pi(\mathcal{X}|\theta)\\
&\propto \pi(\theta)\cdot\prod_{i=2}^{n}\pi(x_i|x_{i-1},\theta)\\
&\propto \pi(\theta)\cdot\prod_{i=2}^{n}\left(\frac{\pi(x_i|x_{i-1},\theta){\color{blue}\pi(\theta)}}{{\color{blue}\pi(\theta)}}\right)\\
&\propto \pi(\theta)^{(2-n)}\prod_{i=2}^{n}\pi(x_i|x_{i-1},\theta)\pi(\theta)\\
&\propto \pi(\theta)^{(2-n)}\prod_{i=2}^{n}\phi_i(\theta).
\end{aligned}
$$

where

- $\phi_i(\theta) = c_i^{-1}\pi(x_i|x_{i-1})\pi(\theta)$
- $c_i = \int \pi(x_i|x_{i-1})\pi(\theta)\,\mathrm{d}\theta$ [normalising constant]

Factorising $\pi(\theta|\mathcal{X}) \to$ PW-EBC/PW-ABC

Essentially, the density of the posterior distribution of interest, $\pi(\theta|\mathcal{X})$, has been decomposed into a product involving densities $\phi_i(\theta)$, each of which depends only on a *pair* of data points $\{x_{i-1}, x_i\}$:

$$\pi(\theta|\mathcal{X}) \propto \pi(\theta)^{(2-n)} \prod_{i=2}^{n} \phi_i(\theta) \tag{2}$$

- If $\pi(x_i|x_{i-1}, \theta)$ is not available/intractable/difficult to compute then so $\phi_i(\theta)$ is and decomposing $\pi(\theta|\mathcal{X})$ will not be of much help.

- However, if we can simulate from each distribution with density $\propto \phi_i(\theta)$, i.e. simulate $x_i|x_{i-1}$, then it turns out that we can recover the posterior density, $\pi(\theta|\mathcal{X})$.

EBC/ABC Within in Each Interval

Although the transition density $\pi(x|x_{i-1})$ might be intractable, we can draw samples from each density

$$\phi_i(\theta) \propto \pi(\theta)\pi(x_i|x_{i-1}, \theta), \qquad i = 2, \ldots n.$$

using the following algorithm:

Algorithm 4 : EBC (ABC) within each interval

1: Sample θ^* from $\pi(\theta)$.
2: Simulate $x_i^\star|x_{i-1}$ from the model using $\theta^\star$.
3: Accept θ^* if $x_i = x_i^\star$ (or $d\big(s(x_i), s(x_i^*)\big) \leq \varepsilon$), otherwise reject.
4: Repeat.

In other words, apply (independent) EBC/ABC for each pair/inteval (x_i, x_{i-1}) to draw from each density $\phi_i(\theta)$.

Putting it altogether . . . → PW-ABC

Algorithm 5 Piece-Wise Approximate Bayesian Computation

for $i = 2$ to n **do**

a: Apply the ABC Algorithm to draw m approximate (or exact, if $s(\cdot) = \mathrm{Identity}(\cdot)$ and $\varepsilon = 0$) samples rom $\tilde{\varphi}_i(\theta)$;

b: Using the samples calculate a density estimate, $\hat{\varphi}_i(\theta)$, of $\tilde{\varphi}_i(\theta)$.

end for

Substitute the density estimates $\hat{\varphi}_i(\theta)$ into (2) to calculate an estimate, $\hat{\pi}(\theta|x)$, of $\pi(\theta|x)$.

(KDE) PW–ABC

- Somehow, we need to derive an estimate of each density using the samples derived in Step 2.
- Such an approach requires a kernel density estimation (KDE) on each $\phi_i(\theta)$ ...
- ... and then multiplying the KDEs pointwise adjusting for the $(n-2)$ prior densities.
- In principle this should work ... and **it does work**, as long as you are careful and you have a descent number of posterior samples in each interval!

(Gaussian) PW–ABC

- KDEs can be hard to deal with; especially in high dimensions!
- Alternatively, we could approximate each $\phi_i(\theta)$ with a (multivariate) Gaussian distribution

$$\widehat{\phi_i(\theta)} = MVN(\mu_i, \Sigma_i)$$

  where μ_i and Σ_i could be the sample mean and the sample variance-covariance matrix;
- Take advantage of the appealing property that the product $\prod_{i=2}^{n} \widehat{\phi_i(\theta)}$ leads to another Gaussian density too . . . ,
- . . . which combined with $(n-2)$ (Gaussian) prior densities leads, finally, to a Gaussian approximation to the full posterior density $\pi(\theta|\mathcal{X})$

(KDE) vs (Gaussian) PW–ABC

- KDEs are known to perform poorly on bounded supports $\rightarrow$ transform the parameters (θ).
- Which Kernel to use?

  We follow Fukunaga (1972) “sphering approach” which selects the bandwidth so that the shape of the kernel mimics the shape of the sample;

- Easy to select an “optimal” bandwidth when doing KDE in each interval, but not so easy when looking at the product of KDEs.
- The Gaussian approximation to each $\phi_i(\theta)$ may not be necessarily good and this will lead to biased estimates $\pi(\theta|\mathcal{X})$.

Applications of PW-EBC/ABC: INAR processes

- Consider the following integer-valued autoregressive model of order 1, known as INAR(1) [Al-Osh and Alzaid, 1987],:

$$X_t = \alpha \circ X_{t-i} + Z_t, \quad t \in \mathbb{Z},$$

  where Z_t are i.i.d. integer-valued random variables and assumed to be independent of the X_t.

- The operator $\alpha\circ$ denotes binomial thinning defined by

$$\alpha \circ W = \begin{cases} \text{Binomial}(W, \alpha), & W > 0, \\ 0, & W = 0, \end{cases}$$

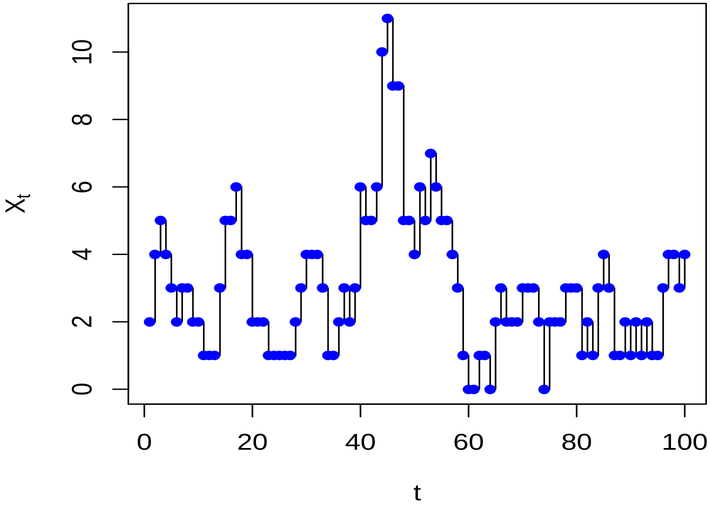
- This model falls into the class of models that one can take advantage of Piecewise approaches.

PW−EBC on INAR(1)

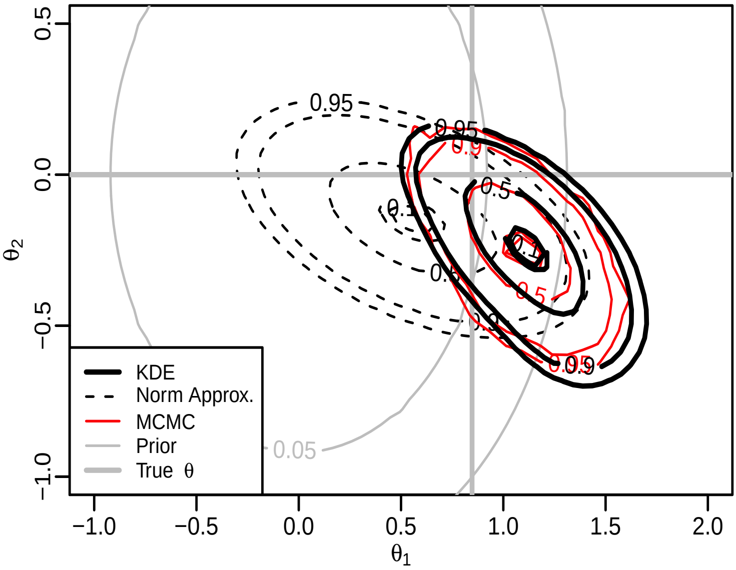
- We generated 100 observations from an INAR(1) process using parameters $\theta = (\alpha, \lambda) = (0.7, 1)$ and $X(0) = 2$
- We make inference on the transformed parameters $\widetilde{\alpha} = \mathrm{logit}(\alpha) = \log(\alpha) - \log(1-\alpha)$ and $\widetilde{\lambda} = \log(\lambda)$ ...
- ... with priors of $\mathrm{Norm}(0, 3^2)$ on the transformed parameters.

- For the EBC algorithm (on the whole dataset) the probability of acceptance is around 10^{-100}, which is prohibitively small.
- Even the ABC algorithm requires a value of ϵ so large that sequential approaches are needed, e.g. SMC-ABC, [Toni *et* al., 2009].

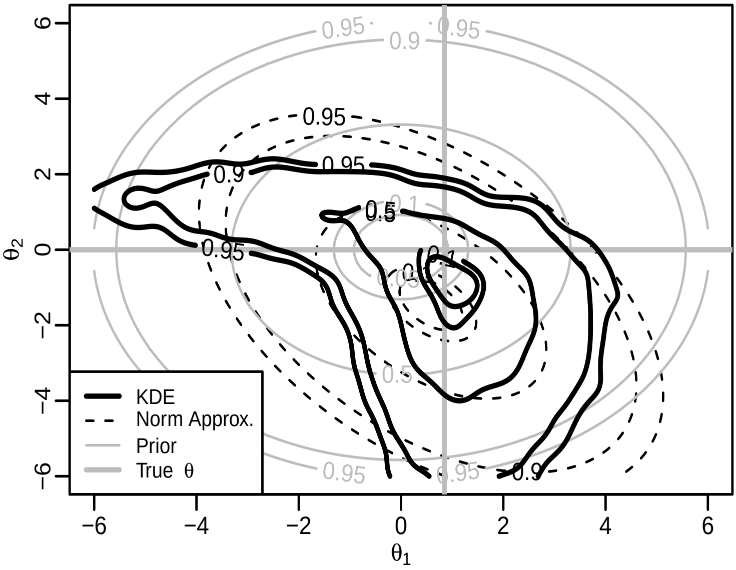
The INAR Dataset



PWEBC on INAR Models



(Gaussian)PW-EBC Does Not Seem to Work



Cox–Ingersoll–Ross Model

The CIR model is a stochastic differential equation (SDE) describing evolution of an interest rate, $X(t)$:

$$\mathrm{d}X(t) = a(b - X(t))\mathrm{d}t + \sigma\sqrt{X(t)}\mathrm{d}W(t),$$

where a, b and σ respectively determine the reversion speed,

long-run value and volatility, and where $W(t)$ denotes a standard Brownian motion.

The density of $X(t_i)|X(t_j), a, b, \sigma$ $(t_i > t_j)$ is a non-central chi-square and hence the likelihood is known in closed form.

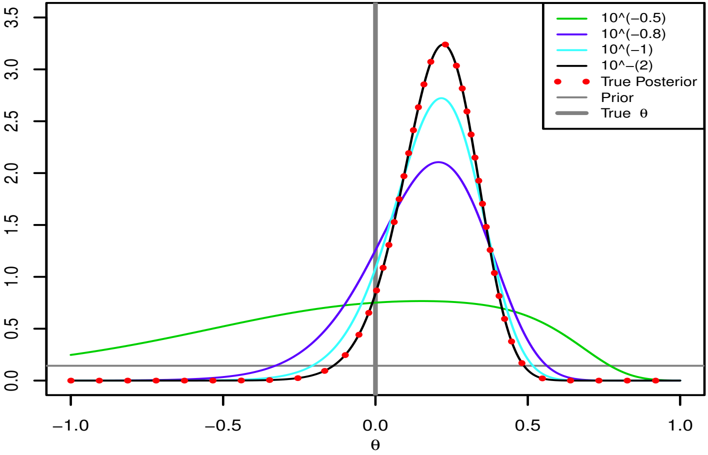
This examples allows to illustrate the use of PW-ABC in the context of continuous data.

Cox–Ingersoll–Ross Model: Some Data

- We generated $n = 10$ equally spaced observations from a CIR process with parameters $(a, b, \sigma) = (0.5, 1, 0.15)$ and $X(0) = 1$ on the interval $t \in [0, 4.5]$.
- Treating a and σ as known, we performed inference on the transformed parameter $\theta = \log(b)$ with a Uniform prior on the interval $(-5, 2)$.
- Using $\varepsilon = 10^{-2}$ we drew samples of size $m = 10,000$ for each $\varphi_i(\theta)$, $i = 1, \ldots, 9$, achieving acceptance rates around 1.5% on average.

Cox–Ingersoll–Ross Model: Inference

The Figure below shows how the posterior density targeted by PW-ABC depends on ε, and how it converges to the true posterior density as $\varepsilon \to 0$.



Stochastic Lotka–Volterra Dynamics

- The stochastic(LV) model is a model of predator–prey dynamics.
- Let Y_1 and Y_2 denote the number of prey and predators respectively, and suppose Y_1 and Y_2 are subject to the following reactions

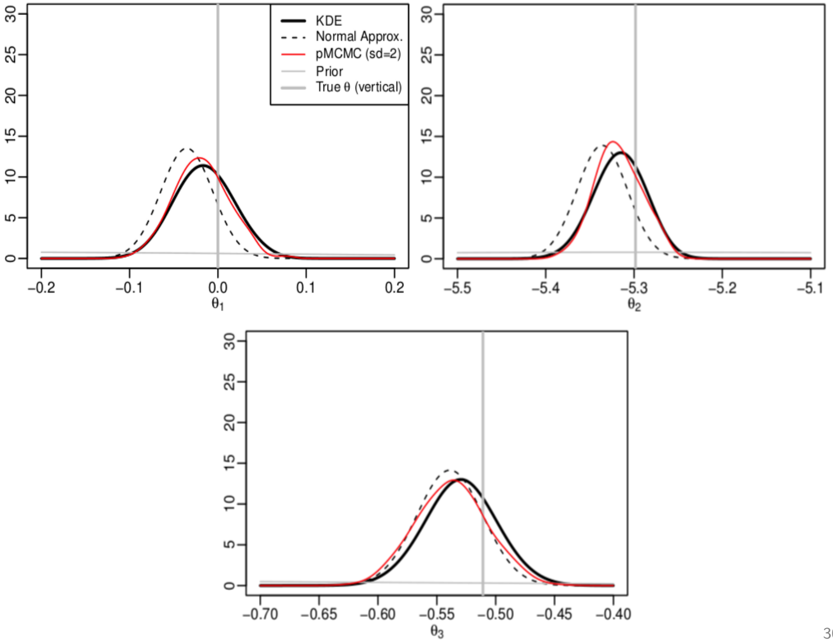
$$Y_1 \xrightarrow{r_1} 2\,Y_1, \qquad Y_1 + Y_2 \xrightarrow{r_2} 2\,Y_2, \qquad Y_2 \xrightarrow{r_3} \emptyset,$$

  which respectively represent prey birth, predation and predator death.
- We wish to make inference for vector of rates $\mathbf{r} = (r_1, r_2, r_3)$.

Likelihood-Based Inference for the LV model

- Inference is simple if the type and precise time of each reaction is observed.
- However, a more common setting is where the population sizes are only observed at discrete time points $\rightarrow$ likelihood is not available.
- Reversible-Jump MCMC has been developed in this context [e.g. Boys *et* al., 2008] but require considerable expertise to implement.
- Other approaches include model approximations using SDEs (e.g. Goligthly and Wilkinson 2006, 2007) and more recently, Particle MCMC (Wilkinson, 2012)
- On the other hand, simulating realizations from this model is straightfoward (e.g. using the Gillespie algorithm).

PW-EBC vs Particle MCMC (sd=2)



Conclusions–Remarks

- If $pi(\theta)$ is too uninformative then PW–EBC/ABC will suffer from the same problems as (standard) EBC/ABC $\rightarrow$ use SMC-EBC/ABC within each interval.
- Use a mixture of Gaussians rather than a single one to approximate the posterior in each interval. For efficiency, use some sort sparsity-induced priors.
- Employ PW–ABC within a Sequential Monte Carlo (aka Particle Filters) framework.
- Scope for the theoretical development on the choice of bandwidth for products of KDES.