Recent Advances in Approximate Bayesian Computation (ABC): Inference and Forecasting

Gael Martin (Monash)

Drawing heavily from work with:

David Frazier (Monash University), Ole Maneesoonthorn (Uni. of Melbourne), Brendan McCabe (Uni. of Liverpool), Christian Robert (Université Paris Dauphine; CREST; Warwick) and Judith Rousseau (Université Paris Dauphine; CREST)

Bayes on the Beach, 2017

Overview

- **Goal:** posterior inference on unknown $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- When the DGP $p(\mathbf{y}|\boldsymbol{\theta})$ is **intractable:**
- **i.e. either** (parts of) the DGP **unavailable** in closed form:
    - Continuous time models (unknown transitions)
    - Gibbs random fields (unknown integrating constant);
    - $\alpha-$stable distributions (density function unavailable)
- **Or** dimension of $\boldsymbol{\theta}$ so large:
    - Coalescent trees
    - Large-scale discrete choice models
- that exploration/marginalization **infeasible** via **exact** methods:
- Can/must resort to **approximate** inference

Approximate Methods

- **Goal** then is to produce **an approximation to** $p(\boldsymbol{\theta}|\mathbf{y})$:
- Approximate Bayesian computation (**ABC**)
- Synthetic Likelihood
- Variational Bayes
- Integrated nested Laplace (INLA)
- **ABC** particularly prominent in genetics, epidemiology, evolutionary biology, ecology
- Where move away from **exact** Bayesian inference *also* motivated by certain features of their problems

Approximate Bayesian Computation (ABC)

- Whilst $p(\mathbf{y}|\boldsymbol{\theta})$ is intractable
- $p(\mathbf{y}|\boldsymbol{\theta})$ (and $p(\boldsymbol{\theta})$) **can be simulated from**
- **ABC** requires **only** this feature
- to produce a **simulation-based estimate** of **an approximation to** $p(\boldsymbol{\theta}|\mathbf{y})$
- (Recent reviews: **Marin et al. 2011; Sisson and Fan, 2011; Robert, 2015; Drovandi, 2017**)

Basic ABC Algorithm - Reiterating!

- Aim is to produce **draws** from an **approximation** to $p(\boldsymbol{\theta}|\mathbf{y})$
- and use draws to **estimate** that **approximation**
- The simplest (**accept/reject**) form of the algorithm:

1. Simulate $(\boldsymbol{\theta}^i)$, $i = 1, 2, ..., N$, from $p(\boldsymbol{\theta})$
2. Simulate **psuedo-data** $\mathbf{z}^i$, $i = 1, 2, ..., N$, from $p(\mathbf{z}|\boldsymbol{\theta}^i)$
3. Select $(\boldsymbol{\theta}^i)$ such that:

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z}^i)\} \leq \varepsilon$$

  - $\boldsymbol{\eta}(.)$ is a (vector) **summary statistic**
  - $d\{.\}$ is a distance criterion
  - the tolerance ε is arbitrarily small

Extensions......

1. **Modification** of the basic algorithm
   1. Using different kernels from the **indicator** kernel:

      $$\mathcal{I}\left[d\{\boldsymbol{\eta}(\mathbf{y}),\boldsymbol{\eta}(\mathbf{z}^i)\}\leq\varepsilon\right]$$

      to give higher weight to those draws, $\boldsymbol{\theta}^i$, that produce $\boldsymbol{\eta}(\mathbf{z}^i)$ close to $\boldsymbol{\eta}(\mathbf{y})$
   2. Inserting MCMC or sequential Monte Carlo (SMC) steps to improve upon taking proposal draws from the **prior**
2. **Adjustment** of the ABC draws via (local) **linear** or **non-linear regression techniques**

- **Beaumont et al., 2002; Marjoram et al., 2003 ; Sisson et al., 2007; Beaumont et al., 2009; Blum, 2010**
- $\Rightarrow$ better simulation-based estimates of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ for a given N and a given $\boldsymbol{\eta}(\mathbf{y})$

Choice of summary statistics?

- **However: the critical aspect of ABC is the choice of $\eta(\mathbf{y})$!**
- And, hence, the very definition of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$!!
- In practice: $\boldsymbol{\eta}(.)$ is not **sufficient** $\Rightarrow$
- i.e. $\boldsymbol{\eta}(.)$ does not reproduce information content of $\mathbf{y}$
- Selected draws (as $\varepsilon \longrightarrow 0$) estimate $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ (**not** $p(\boldsymbol{\theta}|\mathbf{y})$)
- Selection of $\boldsymbol{\eta}(.)$ still an open topic, e.g.
  - **Joyce and Marjoram, 2008; Blum, 2010; Fearnhead and Prangle, 2012**

Choice of summaries via an auxiliary model

- In particular, in the spirit of **indirect inference (II):**
  - **Drovandi et al., 2011; Drovandi et al., 2015; Creel and Kristensen, 2015; Martin, McCabe, Frazier, Maneesoonthorn and Robert, 'Auxiliary Likelihood-Based ABC in State Space Models', 2016**
- think about an **auxiliary model** that **approximates** the true (analytically intractable) model
- With associated likelihood function: $L_a(\mathbf{y};\boldsymbol{\beta})$
- Apply **maximum likelihood** est. to $L_a(\mathbf{y};\boldsymbol{\beta}) \Rightarrow \boldsymbol{\eta}(\mathbf{y}) = \widehat{\boldsymbol{\beta}}$
- $\widehat{\boldsymbol{\beta}}$ **asymptotically sufficient** for $\boldsymbol{\beta}$ in the **auxiliary** model
- If approximating model is 'accurate' enough
  - $\widehat{\boldsymbol{\beta}}$ may be 'close to' being **asym. suff.** for $\boldsymbol{\theta}$ in the true model

Validity of ABC?

- Of late?
- Attention has shifted from ABC as a **practical** tool for estimating an inaccessible $p(\boldsymbol{\theta}|\mathbf{y})$ (via $\widehat{p}(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$)
- To the exploration of its **theoretical asymptotic properties**
- i.e. does **ABC** (as based on **some** choice of $\boldsymbol{\eta}(\mathbf{y})$) do sensible things as the **empirical sample size** T gets bigger?
- i.e. is **ABC valid** as an **inferential** method?

The Asymptotics of ABC

- **Frazier, Martin, Robert and Rousseau, 'Asymptotic Properties of Approximate Bayesian Computation', 2017:**
- Address the following questions:
  1. What is the behaviour of $\Pr(\boldsymbol{\theta} \in A | d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon)$ as $T \to \infty$ **and** $\varepsilon \to 0$
     - For arbitrary $\boldsymbol{\eta}(.)$**?**
     - For $\boldsymbol{\eta}(.)$ extracted from an **auxiliary model**?
  2. Can knowledge of this **asymptotic** behaviour inform our choice of ε, N, for some **finite** T?
- So **actually** addressing a **theoretical** and **practical** question
- (See also **Creel et al., 2015; Li and Fearnhead, 2016a,b; Frazier, Robert and Rousseau, 'Model Misspecification in ABC: Consequences and Diagnostics', 2017**)

Why Care?

Question 1: Asymptotic behaviour of ABC?

- Unless $\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\theta})$ in **exponential family**
- $\boldsymbol{\eta}(\mathbf{y})$ cannot be **sufficient** for $\boldsymbol{\theta}$ and:

$$\Pr(\boldsymbol{\theta} \in A | d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon) \neq \Pr(\boldsymbol{\theta} \in A | \mathbf{y}\}$$

- No real way of quantifying the $\neq$
- Still need some guarantee that our inference is 'valid' in some sense
- Minimum requirement here (surely!) is that:
  - for T 'large enough'
  - the ABC posterior **concentrates** around (true) $\boldsymbol{\theta}_0$:

$$\Pr(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta | d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon) \xrightarrow{p} 0 \textit{ for any } \delta > 0$$

  - i.e. that **Bayesian consistency** holds

Why Care?

Question 1: Asymptotic behaviour of ABC?

- Would also like some guarantee of a sensible limiting shape
    - e.g. **asymptotic normality**
- Plus - heretically - some knowledge of the **asymptotic sampling** distribution of an ABC point estimat**or** (e.g. ABC posterior mean)

Why Care?

Question 2: Choice of tolerance?

- Prevailing wisdom? Take ε as small as possible!
  - $\Rightarrow$ selecting $\boldsymbol{\theta}^{(i)}$ for which $\boldsymbol{\eta}(\mathbf{z}^{(i)}) \approx \boldsymbol{\eta}(\mathbf{y})$
  - $\approx$ represent draws from $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$
- But ABC is costly to implement with small ε
- To maintain a given Monte Carlo error in **estimating** $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ from the **selected** draws
- Need to **increase** N as ε **decreases**!
- But is there a point beyond which taking ε smaller is not helpful?
  - Yes!
  - Related to the conditions required for **asymptotic normality**

Why Care?

- In addition......we **now** know
- (based on more recent explorations.....)
- that the **asymptotic behaviour** of has important ramifications for **forecasting** as based on $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$!
- later......

The Asymptotics of ABC

- **Frazier, Martin, Robert and Rousseau, 'Asymptotic Properties of Approximate Bayesian Computation', 2017:**
- Address three theoretical questions:

1. Does $\Pr(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| > \delta | d\{\boldsymbol{\eta}(\mathbf{y}),\boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon_T) \xrightarrow{p} 0$ for any $\delta > 0$, and for some $\varepsilon_T \to 0$ as $T \to \infty$, for any given $\boldsymbol{\eta}(\mathbf{y})$?
    - i.e. does **Bayesian consistency** hold? **Theorem 1**
2. What is the **asymptotic shape** of (a standardized version of) $\Pr(\boldsymbol{\theta} \in A | d\{\boldsymbol{\eta}(\mathbf{y}),\boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon_T)$, for any given $\boldsymbol{\eta}(\mathbf{y})$
    - i.e. does **asymptotic normality** hold? **Theorem 2**
3. What are the (sampling) properties of the **ABC posterior mean**?
    - Is it **asymptotically normal**? Is it **asy. unbiased**? **Theorem 3**

- What is the required rate $\varepsilon_T \to 0$ for all three results??

Key Assumptions

- Assume:

A1. $\boldsymbol{\eta}(\mathbf{z}) \xrightarrow{P} \mathbf{b}(\boldsymbol{\theta})$ = 'binding function'

A2. Need the presence of **prior mass** near $\mathbf{b}(\boldsymbol{\theta_0})$

A3. The continuity and **injectivity** of $\mathbf{b} : \boldsymbol{\Theta} \to \mathcal{B}$

- i.e. that $\boldsymbol{\theta_0}$ is 'identified' via $\mathbf{b}(\boldsymbol{\theta_0})$

Overview of Key Theoretical Results

- **Theorem 1** :
- Under **A1-A3** have **posterior concentration** for any $\varepsilon_T = o(1)$
- To say something about the **rate of posterior concentration**
  - We require an **additional assumption** on the **tail behaviour of** $\boldsymbol{\eta}(\mathbf{z})$ (around $\mathbf{b}(\boldsymbol{\theta})$)
  - Concentration rate is **faster** the thinner is the (assumed) tail behaviour of $\boldsymbol{\eta}(\mathbf{z})$
  - Concentration rate is **faster** the larger is the (assumed) prior mass near the truth

Overview of Key Theoretical Results

- An arbitrary $\varepsilon_T = o(1)$ **will not** however necessarily yield **asymptotic normality**
- Need a more **stringent** condition on ε_T for the **Gaussian shape**
- **+ need a CLT** for $\boldsymbol{\eta}(\mathbf{z})$
- Assume some *common* (and canonical) rate $\sqrt{T}$ for all elements of $\boldsymbol{\eta}(\mathbf{y})$

Overview of Key Theoretical Results

- **Theorem 2**:
- Given $\varepsilon_T = o(1/\sqrt{T})$:

$$\Pr(\boldsymbol{\theta} \in A | d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon_T) \xrightarrow{p} \Phi(A)$$

- $\Rightarrow$ **asymptotic normality (Bernstein-von Mises)**
- $\Rightarrow$ Bayesian credible intervals will have correct frequentist coverage (asymptotically)
- ($\varepsilon_T = O(1/\sqrt{T})$ yields some shape information but not normality......)

Overview of Key Theoretical Results

- **Theorem 3**
- Does **asy. norm** of ABC posterior mean **require BvM**? **No!**
- For any $\varepsilon_T = o(1)$:

$$E(\boldsymbol{\theta}|d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\} \leq \varepsilon_T) \xrightarrow{d} N$$

- i.e. **asy. norm** of the ABC posterior mean requires no *particular* rate for the tolerance!
- However, require $\varepsilon_T = o(1/T^{0.25})$ for $E(\boldsymbol{\theta}|\ldots)$ to also be **asymptotically unbiased** as an estimator of $\boldsymbol{\theta}_0$
- But even this is a **less stringent** requirement on ε_T than that required for the **BvM** ($\varepsilon_T = o(1/T^{0.5})$)
- $\Rightarrow$ point estimation via 'easier' than acquisition of **BvM**

Role of the Binding Function??

- **Killer** condition (for **all asymptotic results re. θ_0)**:

$$\text{binding function} : \mathbf{b}(\cdot) \text{ is one-to-one in } \boldsymbol{\theta}$$

- Required to **uniquely identify** θ_0 via $\mathbf{b}(\theta_0)$
- Identification hard to achieve in practice!
- Difficult to even verify!
- Why? $\mathbf{b}(\cdot)$ is unknown in closed form (**in practice**)!
- One-to-one condition *also required* for (frequentist) methods of **indirect inference** etc.
- Verification remains an open problem

Practical Implications of Results?

- Standard practice: select draws of $\boldsymbol{\theta}$ that yield distances:

$$d\{\boldsymbol{\eta}(\mathbf{y}), \boldsymbol{\eta}(\mathbf{z})\}$$

  that are less than some α quantile (e.g. $\alpha = 0.01$)
- We link $\varepsilon_T = o(1)$ to $\alpha_T = o(1)$
- e.g: $\varepsilon_T = o(1/\sqrt{T})$ (required for **BvM**)
- $\Leftrightarrow \alpha_T = o(1/\left(\sqrt{T}\right)^{k_\theta})$ ($k_\theta = \dim(\boldsymbol{\theta})$)
- Larger $k_\theta \Rightarrow$ smaller α_T
- If wish to maintain the same Monte Carlo error
- Have to increase N (and, hence computational burden) as T increases
- And even more so, the larger is k_θ!

Practical Implications of Results?

- $k_\eta = \dim(\eta(\mathbf{y}))$ **can exacerbate the problem** once Monte Carlo error is taken into account.
- **Question:** do we gain anything by decreasing ε_T (and hence α_T) below that required for the **BvM??**
- (i.e. the very strictest requirement on ε_T from our theoretical results)
- i.e. can we **cap** the computational burden??
- Cutting to the chase....
- Using a simple example in which $p(\boldsymbol{\theta}|\mathbf{y})$ has closed form
- Find **no gain in accuracy** after $\alpha_T = o(1/\left(\sqrt{T}\right)^{k_\theta})$

Key Messages?

- Link between ABC tolerance (ε_T) and the asymptotic behaviour of ABC is important (and subtle)
- Posterior normality requires a more stringent condition on ε_T
- and, hence, a higher computational burden, than do other asymptotic results
- Rebuke conventional wisdom on choice of ε_T (α_T)
- Care to be taken in choice of summary statistics
- With **injectivity** underpinning all asymptotic results
- **Question remaining?.....**
- What is the impact on **Bayesian forecasting** of using $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ rather than $p(\boldsymbol{\theta}|\mathbf{y})$ to quantify parameter uncertainty?
- And do the **asymptotic properties** of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ matter?

Exact Bayesian Forecasting

- The **Bayesian paradigm:**
- **Quantifying** uncertainty about:

unknown|known

- **using probability**
- In **forecasting**, quantity of interest is y_{T+1};

$$\begin{aligned} p_{exact}(y_{T+1}|\mathbf{y}) &= \int_{\boldsymbol{\theta}} p(y_{T+1}, \boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &= \int_{\theta} p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \\ &= E_{\boldsymbol{\theta}|\mathbf{y}}\left[p(y_{T+1}|\boldsymbol{\theta}, \mathbf{y})\right] \end{aligned}$$

- **Marginal** predictive = expectation of the **conditional** predictive
- **Conditional** predictive reflects the assumed **model**

Exact Bayesian Forecasting

- The expectation is w.r.t: $p(\boldsymbol{\theta}|\mathbf{y})$
- Given M draws from $p(\boldsymbol{\theta}|\mathbf{y})$, $p_{exact}(y_{T+1}|\mathbf{y})$ can be **estimated** as
  1. either:
  $$\widehat{p_{exact}(y_{T+1}|\mathbf{y})} = \frac{1}{M}\sum_{i-1}^{M} p(y_{T+1}|\boldsymbol{\theta}^{(i)}, \mathbf{y})$$
  2. or: $\widehat{p_{exact}(y_{T+1}|\mathbf{y})}$ constructed from draws of $y_{T+1}^{(i)}$ extracted from $p(y_{T+1}|\boldsymbol{\theta}^{(i)}, \mathbf{y})$
- $\Rightarrow$ **exact Bayesian forecasting** (up to simulation error)
- **Note**: while **only 1.** requires $p(y_{T+1}|\boldsymbol{\theta}^{(i)}, \mathbf{y})$ to be available in closed form
- **Both 1. and 2.** require simulation from $p(\boldsymbol{\theta}|\mathbf{y}) \Rightarrow$ (broadly speaking) requires $p(\mathbf{y}|\boldsymbol{\theta})$ to be available

Approximate Bayesian Forecasting

- **Frazier, Maneesoonthorn, Martin and McCabe, 'Approximate Bayesian Forecasting', 2017:**
- How to conduct Bayesian forecasting when the DGP $p(\mathbf{y}|\boldsymbol{\theta})$ is **intractable?**
- And an **approximation to** $p(\boldsymbol{\theta}|\mathbf{y})$ is used to quantify uncertainty about $\boldsymbol{\theta}$?
- $\Rightarrow$ an **approximation to** $p_{exact}(y_{T+1}|\mathbf{y})$
- Focus is on approximating $p(\boldsymbol{\theta}|\mathbf{y})$ via **ABC**
- $\Rightarrow$ Bring insights from **inference** $\Rightarrow$ **forecasting** realm
- No-one has looked at the use of ABC (and the choice of $\boldsymbol{\eta}(\mathbf{y})$) in a **forecasting** context

Approximate Bayesian Forecasting

- ABC automatically yields draws from $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ as the selected draws from the ABC algorithm are used to estimate $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$!
- Hence, we use those selected draws of $\boldsymbol{\theta}$ to estimate:

$$\begin{aligned} p_{ABC}(y_{T+1}|\mathbf{y}) &= \int p(y_{T+1}|\boldsymbol{\theta},\mathbf{y})p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))d\boldsymbol{\theta} \\ &= \text{an ‘}\textbf{approximate Bayesian predictive}\text{’} \end{aligned}$$

- But what is $p_{ABC}(y_{T+1}|\mathbf{y})$??
- Is it a proper predictive density function??
- How does it relate to $p_{exact}(y_{T+1}|\mathbf{y})$??
- We show that $p_{ABC}(y_{T+1}|\mathbf{y})$ **is** a **proper** density function
- But that:

$$p_{ABC}(y_{T+1}|\mathbf{y}) = p_{exact}(y_{T+1}|\mathbf{y}) \text{ iff } \boldsymbol{\eta}(\mathbf{y}) \textbf{ is sufficient}$$

Approximate Bayesian Forecasting

- **Questions!!**
  1. What is the relationship between $p_{exact}(y_{T+1}|\mathbf{y})$ and $p_{ABC}(y_{T+1}|\mathbf{y})$ as $T \to \infty$?
     - What role does **Bayesian consistency** of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$ play here?
  2. How do we **formalize** and **quantify** the loss when we move from $p_{exact}(y_{T+1}|\mathbf{y})$ to $p_{ABC}(y_{T+1}|\mathbf{y})$?
  3. How does one compute $p_{ABC}(y_{T+1}|\mathbf{y})$ in **state space models**?
     - Does one condition **state inference** only on $\boldsymbol{\eta}(\mathbf{y})$?
  4. How should one **choose** $\boldsymbol{\eta}(\mathbf{y})$ in an empirical setting?
     - Why not use **forecasting performance** to determine $\boldsymbol{\eta}(\mathbf{y})$?
- Questions have a theoretical **and** a practical dimension

Q1: Bayes consistency and ‘merging’ of forecasts

- What happens as $T \to \infty$?
- **Blackwell and Dubins (1962)**:
- Two predictive distributions, $P_{\mathbf{y}}$ and $G_{\mathbf{y}}$, ‘**merge**’ if:

$$\rho_{TV}\{P_{\mathbf{y}}, G_{\mathbf{y}}\} = \sup_{B \in \mathcal{F}} |P_{\mathbf{y}}(B) - G_{\mathbf{y}}(B)| = o_{\mathbb{P}}(1)$$

- **Theorem 1:**:
- Under the conditions for the Bayesian consistency of $p(\boldsymbol{\theta}|\mathbf{y})$ and $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$: $P_{exact}(\cdot)$ and $P_{ABC}(\cdot)$ merge
- $\Rightarrow$ for large enough T **exact** and **ABC-based** predictions are **equivalent!**

Q1: Example: MA(2): T = 500

- Consider (simple) example used in **Marin et al., 2011**:

$$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}$$

- $e_t \sim i.i.d.N(0, \sigma_0)$ with **true:** $\theta_{10} = 0.8$; $\theta_{20} = 0.6$; $\sigma_0 = 1.0$
- Use **sample autocovariances**

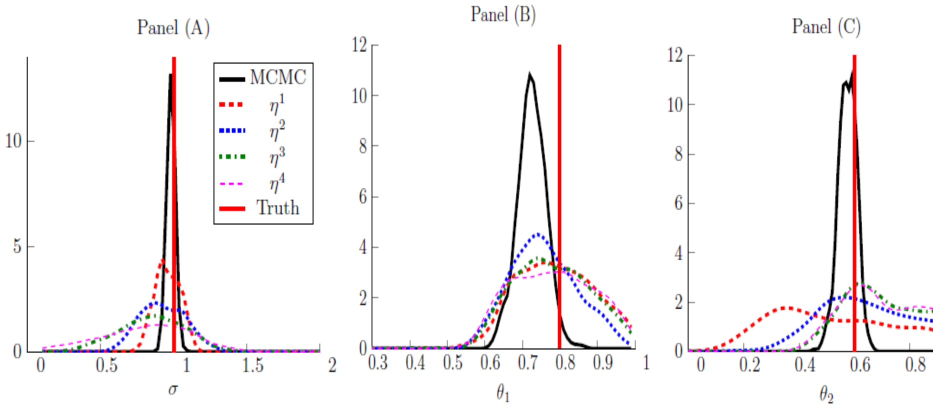
$$\gamma_l = cov(y_t, y_{t-l})$$

- to construct (alternative vectors of) summary statistics:

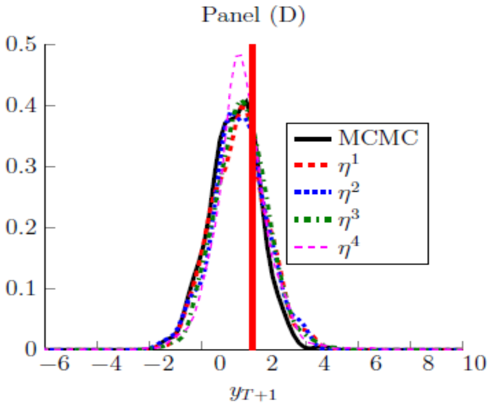
$$\begin{aligned} \eta^{(1)}(\mathbf{y}) &= (\gamma_0, \gamma_1)'; \ \eta^{(2)}(\mathbf{y}) = (\gamma_0, \gamma_1, \gamma_2)' \\ \eta^{(3)}(\mathbf{y}) &= (\gamma_0, \gamma_1, \gamma_2, \gamma_3)'; \ \eta^{(4)}(\mathbf{y}) = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)' \end{aligned}$$

- MA dependence $\Rightarrow$ no reduction to sufficiency possible
  - $\Rightarrow p(\boldsymbol{\theta}|\eta^{(j)}(\mathbf{y})) \neq p(\boldsymbol{\theta}|\mathbf{y})$ for all $j = 1, 2, 3, 4$
  - **What about** $p_{ABC}(y_{T+1}|\mathbf{y})$ *versus* $p_{exact}(y_{T+1}|\mathbf{y})$??

Posterior densities: exact and ABC: T = 500



Predictive densities: exact and ABC: T = 500!

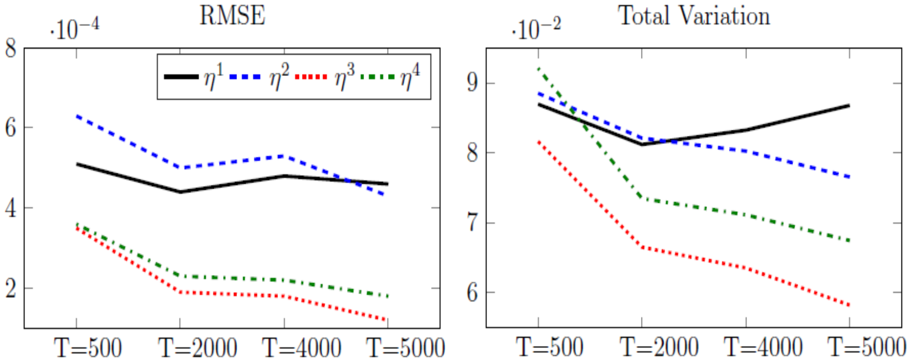


- For large T : the exact and approximate predictives are **very** similar - for all $\eta^{(j)}(\mathbf{y})$!

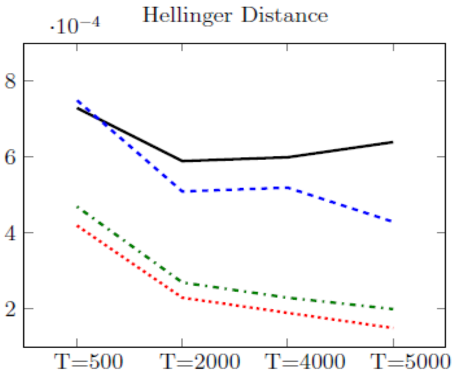
Q1: Example: MA(2); T=500, 2000, 4000, 5000

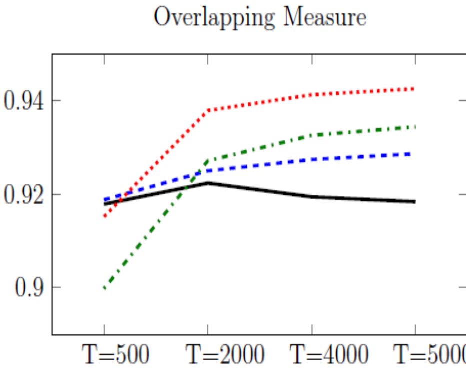
- $\eta^{(1)}(\mathbf{y})$, $\eta^{(2)}(\mathbf{y})$, $\eta^{(3)}(\mathbf{y})$, $\eta^{(4)}(\mathbf{y})$
- $p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(j)}(\mathbf{y}))$ Bayesian consistent for $j = 2, 3, 4$ only
- $\Rightarrow$ expect to see evidence of merging only for $j = 2, 3, 4$
- Measure proximity of $p_{exact}(y_{T+1}|\mathbf{y})$ and $p_{ABC}(y_{T+1}|\mathbf{y})$ using:
    - RMSE of difference between the cdfs ($\downarrow$ as $T\uparrow$)
    - Total variation between the cdfs ($\downarrow$ as $T\uparrow$)
    - Hellinger distance between the cdfs ($\downarrow$ as $T\uparrow$)
    - Degree of overlap between the pdfs ($\uparrow$ as $T\uparrow$)
- All averaged over 100 replications of $\mathbf{y}$

Q1: Example: MA(2); T=500, 2000, 4000, 5000



Q1: Example: MA(2); T=500, 2000, 4000, 5000





- Bayesian consistency in action!

Q2: Quantifying Loss of Accuracy?

- In summary:
  - Under Bayes consistency, $p_{ABC}(y_{T+1}|\mathbf{y})$ and $p_{exact}(y_{T+1}|\mathbf{y})$ equivalent for $T \to \infty$
  - Even for finite T (and lack of consistency) little difference discerned....
- Can we **quantify** accuracy loss?
- Let $S(p_{exact}, y_{T+1})$ be a proper scoring rule (e.g. the log score)
- Define **expected score** under the **truth:**

$$\mathbb{M}(p_{exact}, p_{truth}) = \int_{y\in\Omega} S(p_{exact}, y_{T+1}) \underbrace{p(y_{T+1}|\boldsymbol{\theta}_0, \mathbf{y})}_{p_{truth}} dy_{T+1}$$

Q2: Quantifying Loss of Accuracy?

- **Theorem 2:**: Under Bayes consistency for $p(\boldsymbol{\theta}|\mathbf{y})$ and $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$, if $S(\cdot,\cdot)$ is a strictly proper scoring rule:

1. $|\mathbb{M}(p_{exact}, p_{truth}) - \mathbb{M}(p_{ABC}, p_{truth})| = o_{\mathbb{P}}(1)$;
2. $|\mathbb{E}_{\mathbf{y}}[\mathbb{M}(p_{exact}, p_{truth})] - \mathbb{E}_{\mathbf{y}}[\mathbb{M}(p_{ABC}, p_{truth})]| = o(1)$;
3. 1. and 2. are exactly satisfied if and only if $\eta(\mathbf{y})$ is sufficient for $\mathbf{y}$.

- Either:
    1. **conditionally** (on a given $\mathbf{y}$) or
    2. **unconditionally** (over $\mathbf{y}$)

- For $T \rightarrow \infty$ **approximate forecasting** incurs **no accuracy loss**
    - Other side of the **merging** coin

Q2: Quantifying Loss of Accuracy?

- What if we invoke **more** than Bayes consistency?
- Invoking the (**Cramer Rao**) **efficiency** of the **MLE** (relative to the **ABC** posterior mean):

$$\mathbb{M}(p_{exact}, p_{truth}) \;\geq\; \mathbb{M}(p_{ABC}, p_{truth})$$

$$\mathbb{E}_{\mathbf{y}}\left[\mathbb{M}(p_{exact}, p_{truth})\right] \;\geq\; \mathbb{E}_{\mathbf{y}}\left[\mathbb{M}(p_{ABC}, p_{truth})\right]$$

- $\Rightarrow$ for large (but finite) T would expect the **exact** predictive to yield **higher scores** than the **approximate** predictive**!**

Q2: Example: MA(2): $T = 500$

- **Average predictive scores** over 500 out-of sample values:

	ABC av. score				Exact av. score
	$\eta^{(1)}(\mathbf{y})$	$\eta^{(2)}(\mathbf{y})$	$\eta^{(3)}(\mathbf{y})$	$\eta^{(4)}(\mathbf{y})$	
LS	-1.43	-1.42	-1.43	-1.43	**-1.40**
QS	0.28	0.28	0.28	0.28	**0.29**
CRPS	-0.57	**-0.56**	-0.57	-0.57	**-0.56**

- Loss **is** incurred by being **approximate**
- But it is **negligible!**
- (Including for 'non-consistent' $\eta^{(1)}(\mathbf{y})$)
- Computational gain?
  - $p_{ABC}(y_{T+1}|\mathbf{y})$: 3 seconds
  - $p_{exact}(y_{T+1}|\mathbf{y})$: 360 seconds!

Q3: ABC prediction in state space models?

- **True model** (for financial return, $y_t = \ln P_t - \ln P_{t-1}$), **SV**:

$$\begin{aligned} y_t &= \sqrt{V_t}\varepsilon_t; \qquad \varepsilon_t \sim i.i.d.N(0,1) \\ \ln V_t &= \theta_1 \ln V_{t-1} + \eta_t; \qquad \eta_t \sim i.i.d.N(0,\theta_2) \end{aligned}$$

- $\boldsymbol{\theta} = (\theta_1, \theta_2)'$
- **Auxiliary model, GARCH:**

$$\begin{aligned} y_t &= \sqrt{V_t}\varepsilon_t; \qquad \varepsilon_t \sim i.i.d.N(0,1) \\ V_t &= \beta_1 + \beta_2 V_{t-1} + \beta_3 y_{t-1}^2 \end{aligned}$$

- Closed form for **auxiliary likelihood** $\Rightarrow \widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)'$
- $\Rightarrow \boldsymbol{\eta}(\mathbf{y})$ and $\boldsymbol{\eta}(\mathbf{z})$

Q3: ABC prediction in state space models?

- **Exact:**

$$p_{exact}(y_{T+1}|\mathbf{y}) = \int_{V_{T+1}} \int_{\mathbf{V}} \int_{\boldsymbol{\theta}} p(y_{T+1}|V_{T+1})$$
$$\times \; p(V_{T+1}|V_T, \boldsymbol{\theta}, \mathbf{y}) \underbrace{p(\mathbf{V}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y})}_{p(\mathbf{V}, \boldsymbol{\theta}|\mathbf{y})} d\boldsymbol{\theta} d\mathbf{V} dV_{T+1}$$

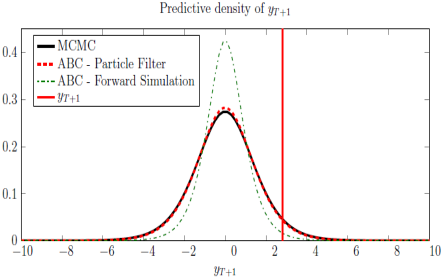
- **MCMC** used to draw from $p(\mathbf{V}, \boldsymbol{\theta}|\mathbf{y})$
- $\Rightarrow$ **independent** draws from $p(V_{T+1}|V_T, \boldsymbol{\theta}, \mathbf{y})$ and $p(y_{T+1}|V_{T+1})$
- $\Rightarrow \widehat{p}_{exact}(y_{T+1}|\mathbf{y})$

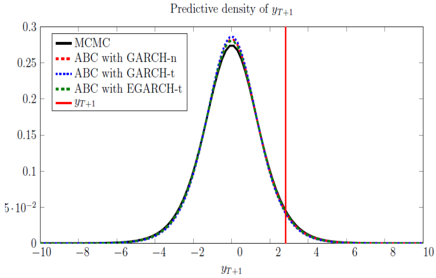
Q3: ABC prediction in state space models?

- **ABC:**

$$p_{ABC}(y_{T+1}|\mathbf{y}) = \int_{V_{T+1}} \int_{\mathbf{V}} \int_{\boldsymbol{\theta}} p(y_{T+1}|V_{T+1})$$
$$\times\ p(V_{T+1}|V_T, \boldsymbol{\theta}, \mathbf{y}) p(\mathbf{V}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) d\boldsymbol{\theta} d\mathbf{V} dV_{T+1}$$

- **ABC** used to draw from $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y}))$
- $\Rightarrow$ **particle filtering** used to integrate out $\mathbf{V}$
- $\Rightarrow$ yields **full posterior inference** (i.e. $|\mathbf{y}$) on V_T
- Exact inference (MCMC) on $\mathbf{V}_{1:T-1}$ not required





- Nature of ABC inference on $\boldsymbol{\theta}$ of little importance.....
- $\Rightarrow$ **All** $p_{ABC}(y_{T+1}|\mathbf{y}) \approx p_{exact}(y_{T+1}|\mathbf{y})$!
- What if condition V_T on $\boldsymbol{\eta}(\mathbf{y})$ only? i.e. omit the **PF** step? Inaccuracy!
- Need to get the predictive **model**: $p(y_{T+1}|V_{T+1})$ and $p(V_{T+1}|V_T, \boldsymbol{\theta}, \mathbf{y})$ 'right'!

Q4: Empirical setting??

- Now to the hard bit......
- Thus far? Have assumed:
  1. That the **DGP:** $p(y_{T+1}, \mathbf{y}, \boldsymbol{\theta}) = p(y_{T+1}|\mathbf{y}, \boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$ is **correct**
  2. That we have access to $p(\boldsymbol{\theta}|\mathbf{y}) \Rightarrow p_{exact}(y_{T+1}|\mathbf{y})$
     - for assessment of $p(\boldsymbol{\theta}|\boldsymbol{\eta}(\mathbf{y})) \Rightarrow p_{ABC}(y_{T+1}|\mathbf{y})$
- In a realistic empirical setting:
  1. We don't know the true **DGP**!!
  2. We are accessing $p_{ABC}(y_{T+1}|\mathbf{y})$ because we cannot (or it is too computationally burdensome) to access $p_{exact}(y_{T+1}|\mathbf{y})$!
  3. $\Rightarrow$ **no benchmark** for $p_{ABC}(y_{T+1}|\mathbf{y})$

Q4: Empirical setting??

- What we CAN access though is **observed** y_{T+1} in a hold out sample
- $\Rightarrow$ if forecasting is the primary aim
- Why not choose $\eta(\mathbf{y})$ (and, hence, $p_{ABC}(y_{T+1}|\mathbf{y})$) according to **actual predictive performance**?

SV model with dynamic jumps and alpha stable errors

- **Two** measurement equations:

$$
\begin{aligned}
r_t &= \exp\left(\frac{h_t}{2}\right)\varepsilon_t + \Delta N_t Z_t;\ \varepsilon_t \sim N(0,1) \\
\ln BV_t &= \psi_0 + \psi_1 h_t + \sigma_{BV}\zeta_t
\end{aligned}
$$

- **Three** state equations:

$$
\begin{aligned}
h_t = \omega + \rho h_{t-1} + \sigma_h \eta_t;\ \eta_t \sim \mathcal{S}(\alpha, -1, 0) \\
Z_t \sim N(\mu, \sigma_z^2) \\
Pr(\Delta N_t = 1 | \mathcal{F}_{t-1}) = \delta_t = \delta + \beta\delta_{t-1} + \gamma\Delta N_{t-1}\ \text{(Hawkes)}
\end{aligned}
$$

- $\Rightarrow$ no closed-form solution for $p(h_t | h_{t-1})$
- $\Rightarrow$ run with **ABC** and **approximate Bayesian forecasting......**

- Choose $\eta(\mathbf{y})$ via **four** different GARCH-type **auxiliary** models **supplemented** with various statistics computed from high-frequency measures of **volatility** and **jumps**
- Compute average scores (for r_t and $\ln BV_t$) and over hold out sample of one trading year:

		Auxiliary model			
		GARCH-N	GARCH-T	TARCH-T	RGARCH
r_t	**LS**	-1.571	-1.280	**-1.202**	-1.945
	QS	0.377	0.474	**0.515**	0.274
	CRPS	-1.515	-1.052	**-0.989**	-2.103
$\ln BV_t$	**LS**	**-2.732**	-2.757	-2.928	-2.827
	QS	**0.095**	0.049	0.016	0.094
	CRPS	-2.038	-1.416	**-1.377**	-2.570

- TGARCH with Student t errors, and various add-ons, the best overall!
- Uniformly so with predicting returns

To come.....

- Questions remain though regarding the **theoretical (asymptotic)** properties of $p_{ABC}(y_{T+1}|\mathbf{y})$ built from such a choice of $\eta(\mathbf{y})$
- Bayesian consistency of $p(\boldsymbol{\theta}|\eta(\mathbf{y}))$ no longer sought
- $\Rightarrow$ merging of $p_{ABC}(y_{T+1}|\mathbf{y})$ and $p_{exact}(y_{T+1}|\mathbf{y})$ no longer an automatic outcome
- However, under **correct model specification:** has been shown to provide an upper bound on the accuracy of $p_{ABC}(y_{T+1}|\mathbf{y})$
- $\Rightarrow$ choosing $\eta(\mathbf{y}) \Rightarrow$ most accurate $p_{ABC}(y_{T+1}|\mathbf{y})$
- $\equiv$ choosing $p_{ABC}(y_{T+1}|\mathbf{y})$ that is closest to $p_{exact}(y_{T+1}|\mathbf{y})$

To come.....

- Under **mis-specification??**
- Still makes perfect sense to pick the $p_{ABC}(y_{T+1}|\mathbf{y})$ with the best forecasting performance!
- What is unclear though is the relationship between $p_{exact}(y_{T+1}|\mathbf{y})$ and $p_{ABC}(y_{T+1}|\mathbf{y})$
- Indeed, in what sense does $p_{exact}(y_{T+1}|\mathbf{y})$ remain **preferable** to $p_{ABC}(y_{T+1}|\mathbf{y})$?
- Are there ways of producing **approximate predictives** that are **robust** to mis-specification?
- ......For another day......