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 $\theta$:

  \[ p(\theta|y) \propto p(y|\theta)p(\theta) \]

- When the DGP $p(y|\theta)$ 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 $\theta$ 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 $p(\theta|y)$:
  - 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

Drawing heavily from work with: Gael Martin (Monash), 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)
Approximate Bayesian Computation (ABC)

- Whilst $p(y|\theta)$ is intractable
- $p(y|\theta)$ (and $p(\theta)$) can be simulated from
- ABC requires only this feature
- to produce a simulation-based estimate of an approximation to $p(\theta|y)$

(Recent reviews: Marin et al. 2011; Sisson and Fan, 2011; Robert, 2015; Drovandi, 2017)
Aim is to produce **draws** from an **approximation** to $p(\theta|y)$ and use draws to **estimate** that **approximation**.

The simplest (**accept/reject**) form of the algorithm:

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

   $$d\{\eta(y), \eta(z^i)\} \leq \epsilon$$

   - $\eta(.)$ is a (vector) **summary statistic**
   - $d\{.\}$ is a distance criterion
   - the tolerance $\epsilon$ is arbitrarily small
1. **Modification** of the basic algorithm
   - Using different kernels from the **indicator** kernel:
     \[
     
     \mathcal{I} \left[ d \{ \eta(y), \eta(z^i) \} \leq \varepsilon \right]
     \]
     to give higher weight to those draws, \( \theta^i \), that produce \( \eta(z^i) \) close to \( \eta(y) \)
   - 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 \( p(\theta|\eta(y)) \) for a given \( N \) and a given \( \eta(y) \)
Choice of summary statistics?

- However: the critical aspect of ABC is the choice of $\eta(y)$!
- And, hence, the very definition of $p(\theta|\eta(y))$!
- In practice: $\eta(.)$ is not sufficient $\Rightarrow$
- i.e. $\eta(.)$ does not reproduce information content of $y$
- Selected draws (as $\varepsilon \to 0$) estimate $p(\theta|\eta(y))$ (not $p(\theta|y)$)
- Selection of $\eta(.)$ 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: \( L_a(y; \beta) \)

- Apply **maximum likelihood** est. to \( L_a(y; \beta) \Rightarrow \eta(y) = \hat{\beta} \)

- \( \hat{\beta} \) **asymptotically sufficient** for \( \beta \) in the **auxiliary** model

- If approximating model is ‘accurate’ enough
  - \( \hat{\beta} \) may be ‘close to’ being **asym. suff.** for \( \theta \) in the true model

Gael Martin (Monash), Drawing heavily from: Recent Advances in Approximate Bayesian Cc
Validity of ABC?

- Of late?

- Attention has shifted from ABC as a **practical** tool for estimating an inaccessible $p(\theta|y)$ (via $\hat{p}(\theta|\eta(y))$)

- To the exploration of its **theoretical asymptotic properties**

- i.e. does **ABC** (as based on **some** choice of $\eta(y)$) do sensible things as the **empirical sample size** $T$ gets bigger?

- i.e. is **ABC** **valid** as an **inferential** method?
Frazier, Martin, Robert and Rousseau, ‘Asymptotic Properties of Approximate Bayesian Computation’, 2017:

Address the following questions:

1. What is the behaviour of $\Pr(\theta \in A|d\{\eta(y), \eta(z)\} \leq \varepsilon)$ as $T \to \infty$ and $\varepsilon \to 0$?
   - For arbitrary $\eta(.)$?
   - For $\eta(.)$ extracted from an auxiliary model?

2. Can knowledge of this asymptotic behaviour inform our choice of $\varepsilon$, $N$, 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 $y \sim p(y|\theta)$ in exponential family
- $\eta(y)$ cannot be sufficient for $\theta$ and:

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

- 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) $\theta_0$:

$$\Pr(\|\theta - \theta_0\| > \delta|d\{\eta(y), \eta(z)\} \leq \varepsilon) \xrightarrow{P} 0 \text{ for any } \delta > 0$$

- 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 estimator (e.g. ABC posterior mean)
Why Care?

Question 2: Choice of tolerance?

- Prevailing wisdom? Take $\varepsilon$ as small as possible!
  - $\Rightarrow$ selecting $\theta^{(i)}$ for which $\eta(z^{(i)}) \approx \eta(y)$
  - $\approx$ represent draws from $p(\theta|\eta(y))$

- But ABC is costly to implement with small $\varepsilon$

- To maintain a given Monte Carlo error in estimating $p(\theta|\eta(y))$ from the selected draws

- Need to increase $N$ as $\varepsilon$ decreases!

- But is there a point beyond which taking $\varepsilon$ smaller is not helpful?
  - Yes!
  - Related to the conditions required for asymptotic normality
In addition......we now know

(based on more recent explorations.....)

that the **asymptotic behaviour** of has important ramifications for **forecasting** as based on \( p(\theta | \eta(y)) \)!

later......
The Asymptotics of ABC

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

1. Does \( \Pr(\|\theta - \theta_0\| > \delta | d\{\eta(y), \eta(z)\} \leq \varepsilon_T) \xrightarrow{p} 0 \) for any \( \delta > 0 \), and for some \( \varepsilon_T \rightarrow 0 \) as \( T \rightarrow \infty \), for any given \( \eta(y) \)?
   - i.e. does Bayesian consistency hold? **Theorem 1**

2. What is the **asymptotic shape** of (a standardized version of) \( \Pr(\theta \in A | d\{\eta(y), \eta(z)\} \leq \varepsilon_T), \) for any given \( \eta(y) \)?
   - 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 \( \varepsilon_T \rightarrow 0 \) for all three results??
Key Assumptions

- Assume:

A1. $\eta(z) \xrightarrow{P} b(\theta) = \text{‘binding function’}$

A2. Need the presence of prior mass near $b(\theta_0)$

A3. The continuity and injectivity of $b : \Theta \rightarrow \mathcal{B}$

- i.e. that $\theta_0$ is ‘identified’ via $b(\theta_0)$
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 \( \eta(z) \) (around \( b(\theta) \))
- Concentration rate is faster the thinner is the (assumed) tail behaviour of \( \eta(z) \)
- Concentration rate is faster the larger is the (assumed) prior mass near the truth
Overview of Key Theoretical Results

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

- **Theorem 2:**

  Given \( \varepsilon_T = o\left(\frac{1}{\sqrt{T}}\right) \):

  \[
  \Pr(\theta \in A \mid d\{\eta(y), \eta(z)\} \leq \varepsilon_T) \xrightarrow{p} \Phi(A)
  \]

  \( \Rightarrow \) **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 \textit{asy. norm} of ABC posterior mean \textit{require BvM}? \textbf{No!}

• For any $\varepsilon_T = o(1)$:

\[
E(\theta | d\{\eta(y), \eta(z)\} \leq \varepsilon_T) \Rightarrow N
\]

• i.e. \textit{asy. norm} of the ABC posterior mean requires no \textit{particular} rate for the tolerance!

• However, require $\varepsilon_T = o(1/ T^{0.25})$ for $E(\theta | ...)$ to also be \textit{asymptotically unbiased} as an estimator of $\theta_0$

• But even this is a \textbf{less stringent} requirement on $\varepsilon_T$ than that required for the \textbf{BvM} ($\varepsilon_T = o(1/ T^{0.5})$)

• $\Rightarrow$ point estimation via ‘easier’ than acquisition of \textbf{BvM}
Role of the Binding Function??

- **Killer** condition (for all asymptotic results re. $\theta_0$):

  
  binding function : $b(\cdot)$ is one-to-one in $\theta$

- Required to **uniquely identify** $\theta_0$ via $b(\theta_0)$

- Identification hard to achieve in practice!

- Difficult to even verify!

- Why? $b(\cdot)$ 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

Gael Martin (Monash), Drawing heavily from Recent Advances in Approximate Bayesian Cc
Practical Implications of Results?

- Standard practice: select draws of $\theta$ that yield distances:
  \[ d\{\eta(y), \eta(z)\} \]
  that are less than some $\alpha$ 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)

- $\iff \alpha_T = o(1/\left(\sqrt{T}\right)^{k_\theta})$ ($k_\theta = \text{dim}(\theta)$)

- Larger $k_\theta \Rightarrow$ smaller $\alpha_T$

- 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 $k_\theta$!
Practical Implications of Results?

- $k_\eta = \text{dim}(\eta(y))$ can exacerbate the problem once Monte Carlo error is taken into account.

- **Question:** do we gain anything by decreasing $\varepsilon_T$ (and hence $\alpha_T$) below that required for the BvM??

- (i.e. the very strictest requirement on $\varepsilon_T$ from our theoretical results)

- i.e. can we **cap** the computational burden??

- Cutting to the chase....

- Using a simple example in which $p(\theta|y)$ 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 ($\varepsilon_T$) and the asymptotic behaviour of ABC is important (and subtle)
- Posterior normality requires a more stringent condition on $\varepsilon_T$ and, hence, a higher computational burden, than do other asymptotic results
- Rebuke conventional wisdom on choice of $\varepsilon_T$ ($\alpha_T$)
- 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 $p(\theta | \eta(y))$ rather than $p(\theta | y)$ to quantify parameter uncertainty?
- And do the asymptotic properties of $p(\theta | \eta(y))$ matter?
Exact Bayesian Forecasting

- The **Bayesian paradigm**: Quantifying uncertainty about: unknown | known

- using probability

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

$$p_{\text{exact}}(y_{T+1}|y) = \int_\theta p(y_{T+1}, \theta|y) d\theta$$

$$= \int_\theta p(y_{T+1}|\theta, y)p(\theta|y) d\theta$$

$$= E_{\theta|y} [p(y_{T+1}|\theta, y)]$$

- **Marginal** predictive = expectation of the **conditional** predictive

- **Conditional** predictive reflects the assumed **model**
Exact Bayesian Forecasting

- The expectation is w.r.t: $p(\theta|y)$

- Given $M$ draws from $p(\theta|y)$, $p_{\text{exact}}(y_{T+1}|y)$ can be estimated as
  - either:
    $$p_{\text{exact}}(y_{T+1}|y) = \frac{1}{M} \sum_{i=1}^{M} p(y_{T+1}|\theta^{(i)}, y)$$
  - or: $p_{\text{exact}}(y_{T+1}|y)$ constructed from draws of $y^{(i)}_{T+1}$ extracted from $p(y_{T+1}|\theta^{(i)}, y)$

- $\Rightarrow$ exact Bayesian forecasting (up to simulation error)

- **Note:** while only 1. requires $p(y_{T+1}|\theta^{(i)}, y)$ to be available in closed form

- **Both 1. and 2.** require simulation from $p(\theta|y) \Rightarrow$ (broadly speaking) requires $p(y|\theta)$ to be available
Approximate Bayesian Forecasting

- **Frazier, Maneesoonthorn, Martin and McCabe, ‘Approximate Bayesian Forecasting’, 2017:**

  How to conduct Bayesian forecasting when the DGP $p(y|\theta)$ is intractable?

- And an **approximation to** $p(\theta|y)$ is used to quantify uncertainty about $\theta$?

  $\Rightarrow$ an **approximation to** $p_{\text{exact}}(y_{T+1}|y)$

- Focus is on approximating $p(\theta|y)$ via ABC

  $\Rightarrow$ Bring insights from **inference** $\Rightarrow$ **forecasting** realm

- No-one has looked at the use of ABC (and the choice of $\eta(y)$) in a **forecasting** context
Approximate Bayesian Forecasting

- ABC automatically yields draws from $p(\theta | \eta(y))$ as the selected draws from the ABC algorithm are used to estimate $p(\theta | \eta(y))$.

- Hence, we use those selected draws of $\theta$ to estimate:
  \[
p_{ABC}(y_{T+1} | y) = \int p(y_{T+1} | \theta, y) p(\theta | \eta(y)) d\theta
  = \text{an ‘approximate Bayesian predictive’}
  \]

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

We show that $p_{ABC}(y_{T+1} | y)$ is a proper density function.

But that:
\[
p_{ABC}(y_{T+1} | y) = p_{\text{exact}}(y_{T+1} | y) \text{ iff } \eta(y) \text{ is sufficient}
\]
Questions!!

1. What is the relationship between $p_{exact}(y_{T+1}|y)$ and $p_{ABC}(y_{T+1}|y)$ as $T \to \infty$?
   - What role does Bayesian consistency of $p(\theta|\eta(y))$ play here?

2. How do we formalize and quantify the loss when we move from $p_{exact}(y_{T+1}|y)$ to $p_{ABC}(y_{T+1}|y)$?

3. How does one compute $p_{ABC}(y_{T+1}|y)$ in state space models?
   - Does one condition state inference only on $\eta(y)$?

4. How should one choose $\eta(y)$ in an empirical setting?
   - Why not use forecasting performance to determine $\eta(y)$?

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, $P_y$ and $G_y$, ‘merge’ if:

$$\rho_{TV}\{P_y, G_y\} = \sup_{B \in \mathcal{F}} |P_y(B) - G_y(B)| = o_{\mathbb{P}}(1)$$

**Theorem 1:**

- Under the conditions for the Bayesian consistency of $p(\theta|y)$ and $p(\theta|\eta(y))$: $P_{\text{exact}}(\cdot)$ and $P_{\text{ABC}}(\cdot)$ merge

$\implies$ 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
  \[ \gamma_l = \text{cov}(y_t, y_{t-l}) \]
  
  - to construct (alternative vectors of) summary statistics:
    \[ \eta^{(1)}(y) = (\gamma_0, \gamma_1)' \quad \eta^{(2)}(y) = (\gamma_0, \gamma_1, \gamma_2)' \]
    \[ \eta^{(3)}(y) = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)' \quad \eta^{(4)}(y) = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)' \]
- MA dependence $\Rightarrow$ no reduction to sufficiency possible
  \[ \Rightarrow p(\theta|\eta^{(j)}(y)) \neq p(\theta|y) \text{ for all } j = 1, 2, 3, 4 \]
- **What about** $p_{ABC}(y_{T+1}|y)$ versus $p_{exact}(y_{T+1}|y)$?
Posterior densities: exact and ABC: $T = 500$

Panel (A)

Panel (B)

Panel (C)
For large $T$ : the exact and approximate predictives are very similar - for all $\eta^{(j)}(y)$!
Q1: Example: MA(2); T=500, 2000, 4000, 5000

- $\eta^{(1)}(y)$, $\eta^{(2)}(y)$, $\eta^{(3)}(y)$, $\eta^{(4)}(y)$

- $\rho(\theta|\eta^{(j)}(y))$ Bayesian consistent for $j = 2, 3, 4$ only

- $\Rightarrow$ expect to see evidence of merging only for $j = 2, 3, 4$

- Measure proximity of $p_{exact}(y_{T+1}|y)$ and $p_{ABC}(y_{T+1}|y)$ 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 $y$
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, $p_{ABC}(y_{T+1}|y)$ and $p_{exact}(y_{T+1}|y)$ 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**:

$$\mathbb{M}(p_{exact}, p_{truth}) = \int_{y \in \Omega} S(p_{exact}, y_{T+1}) p(y_{T+1}|\theta_0, y) dy_{T+1}$$
Q2: Quantifying Loss of Accuracy?

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

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

Either:

1. **conditionally** (on a given \( y \)) or
2. **unconditionally** (over \( y \))

For \( T \to \infty \) **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):

  \[ M(p_{exact}, p_{truth}) \geq M(p_{ABC}, p_{truth}) \]

  \[ E_y [M(p_{exact}, p_{truth})] \geq E_y [M(p_{ABC}, p_{truth})] \]

- ⇒ 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:

<table>
<thead>
<tr>
<th></th>
<th>$\eta^{(1)}(y)$</th>
<th>$\eta^{(2)}(y)$</th>
<th>$\eta^{(3)}(y)$</th>
<th>$\eta^{(4)}(y)$</th>
<th>Exact av. score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS</td>
<td>-1.43</td>
<td>-1.42</td>
<td>-1.43</td>
<td>-1.43</td>
<td>-1.40</td>
</tr>
<tr>
<td>QS</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>CRPS</td>
<td>-0.57</td>
<td>-0.56</td>
<td>-0.57</td>
<td>-0.57</td>
<td>-0.56</td>
</tr>
</tbody>
</table>

- Loss is incurred by being **approximate**
- But it is **negligible!**
- (Including for ‘non-consistent’ $\eta^{(1)}(y)$)
- Computational gain?
  - $p_{ABC}(y_{T+1}|y)$: 3 seconds
  - $p_{exact}(y_{T+1}|y)$: 360 seconds!
Q3: ABC prediction in state space models?

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

  \[
  y_t = \sqrt{V_t} \varepsilon_t; \quad \varepsilon_t \sim i.i.d. N(0, 1)
  \]

  \[
  \ln V_t = \theta_1 \ln V_{t-1} + \eta_t; \quad \eta_t \sim i.i.d. N(0, \theta_2)
  \]

  \( \theta = (\theta_1, \theta_2)' \)

- **Auxiliary model, GARCH:**

  \[
  y_t = \sqrt{V_t} \varepsilon_t; \quad \varepsilon_t \sim i.i.d. N(0, 1)
  \]

  \[
  V_t = \beta_1 + \beta_2 V_{t-1} + \beta_3 y_{t-1}^2
  \]

  Closed form for **auxiliary likelihood** \( \Rightarrow \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)' \)

  \( \Rightarrow \eta(y) \) and \( \eta(z) \)
Q3: ABC prediction in state space models?

- **Exact:**

\[
p_{\text{exact}}(y_{T+1}|y) = \int_{V_{T+1}} \int_{V} \int_{\theta} p(y_{T+1}|V_{T+1}) \\
\times p(V_{T+1}|V,\theta,y) p(V|\theta,y)p(\theta|y) d\theta dV dV_{T+1} \\
\]

- MCMC used to draw from \(p(V,\theta|y)\)

- \(\Rightarrow\) **independent** draws from \(p(V_{T+1}|V,\theta,y)\) and \(p(y_{T+1}|V_{T+1})\)

- \(\Rightarrow \hat{p}_{\text{exact}}(y_{T+1}|y)\)

**Recent Advances in Approximate Bayesian Computation (ABC): Inference and Forecasting**
Q3: ABC prediction in state space models?

- **ABC:**

\[
p_{\text{ABC}}(y_{T+1}|y) = \int_{V_{T+1}} \int_{V} \int_{\theta} p(y_{T+1}|V_{T+1}) \\
\times p(V_{T+1}|V_{T}, \theta, y)p(V|\theta, y)p(\theta|\eta(y)) \, d\theta \, dV \, dV_{T+1}
\]

- **ABC used to draw from** \(p(\theta|\eta(y))\)

- \(\Rightarrow\) **particle filtering** used to integrate out \(V\)

- \(\Rightarrow\) **yields full posterior inference** (i.e. \(\theta \mid y\)) on \(V_{T}\)

- Exact inference (MCMC) on \(V_{1:T-1}\) not required
Nature of ABC inference on $\theta$ of little importance.....

$\Rightarrow \textbf{All } p_{ABC}(y_{T+1}|y) \approx p_{\text{exact}}(y_{T+1}|y)$!

What if condition $V_T$ on $\eta(y)$ only? i.e. omit the PF step? Inaccuracy!

Need to get the predictive model: $p(y_{T+1}|V_{T+1})$ and $p(V_{T+1}|V_T, \theta, y)$ ‘right’!
Now to the hard bit......

Thus far? Have assumed:

1. That the **DGP**: \( p(y_{T+1}, y, \theta) = p(y_{T+1} | y, \theta)p(y | \theta) \) is **correct**
2. That we have access to \( p(\theta | y) \Rightarrow p_{exact}(y_{T+1} | y) \)
   - for assessment of \( p(\theta | \eta(y)) \Rightarrow p_{ABC}(y_{T+1} | y) \)

In a realistic empirical setting:

1. We don’t know the true **DGP**!!
2. We are accessing \( p_{ABC}(y_{T+1} | y) \) because we cannot (or it is too computationally burdensome) to access \( p_{exact}(y_{T+1} | y) ! \)
3. \( \Rightarrow \) **no benchmark** for \( p_{ABC}(y_{T+1} | y) \)
Q4: Empirical setting??

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

- **Two** measurement equations:

\[ r_t = \exp\left(\frac{h_t}{2}\right) \varepsilon_t + \Delta N_t Z_t; \quad \varepsilon_t \sim N(0, 1) \]

\[ \ln BV_t = \psi_0 + \psi_1 h_t + \sigma_{BV} \zeta_t \]

- **Three** state equations:

\[ h_t = \omega + \rho h_{t-1} + \sigma_h \eta_t; \quad \eta_t \sim S(\alpha, -1, 0) \]

\[ Z_t \sim N(\mu, \sigma^2_Z) \]

\[ Pr(\Delta N_t = 1|\mathcal{F}_{t-1}) = \delta_t = \delta + \beta \delta_{t-1} + \gamma \Delta N_{t-1} \text{ (Hawkes)} \]

- \( \Rightarrow \) no closed-form solution for \( p(h_t|h_{t-1}) \)

- \( \Rightarrow \) run with ABC and approximate Bayesian forecasting......
- Choose $\eta(y)$ 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:

<table>
<thead>
<tr>
<th></th>
<th>Auxiliary model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GARCH-N</td>
</tr>
<tr>
<td>$r_t$</td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>-1.571</td>
</tr>
<tr>
<td>QS</td>
<td>0.377</td>
</tr>
<tr>
<td>CRPS</td>
<td>-1.515</td>
</tr>
<tr>
<td>$\ln BV_t$</td>
<td></td>
</tr>
<tr>
<td>QS</td>
<td>0.095</td>
</tr>
<tr>
<td>CRPS</td>
<td>-2.038</td>
</tr>
</tbody>
</table>

- TGARCH with Student $t$ errors, and various add-ons, the best overall!
- Uniformly so with predicting returns
Questions remain though regarding the **theoretical** (asymptotic) properties of $p_{ABC}(y_{T+1}|y)$ built from such a choice of $\eta(y)$

Bayesian consistency of $p(\theta|\eta(y))$ no longer sought

$\Rightarrow$ merging of $p_{ABC}(y_{T+1}|y)$ and $p_{exact}(y_{T+1}|y)$ no longer an automatic outcome

However, under **correct model specification**: has been shown to provide an upper bound on the accuracy of $p_{ABC}(y_{T+1}|y)$

$\Rightarrow$ choosing $\eta(y) \Rightarrow$ most accurate $p_{ABC}(y_{T+1}|y)$

$\equiv$ choosing $p_{ABC}(y_{T+1}|y)$ that is closest to $p_{exact}(y_{T+1}|y)$
To come.....

- Under **mis-specification??**

- Still makes perfect sense to pick the \( p_{ABC}(y_{T+1}|y) \) with the best forecasting performance!

- What is unclear though is the relationship between 
  \( p_{\text{exact}}(y_{T+1}|y) \) and \( p_{ABC}(y_{T+1}|y) \)

- Indeed, in what sense does \( p_{\text{exact}}(y_{T+1}|y) \) remain **preferable** to \( p_{ABC}(y_{T+1}|y) \) ?

- Are there ways of producing **approximate predictives** that are **robust** to mis-specification?

- ......For another day......