Happy ABC: Expectation-Propagation for Summary-Less, Likelihood-Free Inference

Nicolas Chopin

CREST (ENSAE)

joint work with Simon Barthelmé (TU Berlin)
Basic ABC

Data: $y^*$, prior $p(\theta)$, model $p(y|\theta)$. Likelihood $p(y|\theta)$ is intractable.

1. Sample $\theta \sim p(\theta)$
2. Sample $y \sim p(y|\theta)$
3. Accept $\theta$ iff $\|s(y) - s(y^*)\| \leq \epsilon$
The previous algorithm targets:

\[ p_{\varepsilon}(\theta|y^*) \propto p(\theta) \int p(y|\theta) 1_{\{\|s(y) - s(y^*)\| \leq \varepsilon\}} \, dy \]

which approximates the true posterior \( p(\theta|y) \). Two levels of approximation:

1. **Non-parametric error**, governed by “bandwidth” \( \varepsilon \);
   \( p_{\varepsilon}(\theta|y^*) \to p(\theta|s(y^*)) \) as \( \varepsilon \to 0 \).

2. **Bias introduced by summary stat. \( s \)**, since
   \( p(\theta|s(y^*)) \neq p(\theta|y^*) \).

Note that \( p(\theta|s(y^*)) \approx p(\theta|y^*) \) may be a reasonable approximation, but \( p(y^*) \) and \( p(s(y^*)) \) have no clear relation: hence **standard ABC cannot reliably approximate the evidence**.
EP-ABC target

Assume that the data $y$ decomposes into $(y_1, \ldots, y_n)$, and consider the ABC approximation:

$$p_\epsilon(\theta|y^*) \propto p(\theta) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^*, \theta) 1_{\|y_i-y_i^*\| \leq \epsilon} dy_i \right\} \quad (1)$$

Standard ABC cannot target this approximate posterior, because the probability that $\|y_i - y_i^*\| \leq \epsilon$ for all $i$ simultaneously is exponentially small w.r.t. $n$. But it does not depend on some summary stats $s$, and $p_\epsilon(\theta|y^*) \to p(\theta|y^*)$ as $\epsilon \to 0$ (one level of approximation).

The EP-ABC algorithm computes a Gaussian approximation of (1).
Noisy ABC interpretation

Note that the EP-ABC target of the previous slide can be interpreted as the correct posterior distribution of a model where the datapoints are corrupted with a $U[-\epsilon, \epsilon]$ noise, following Wilkinson (2008).
EP: an introduction

Introduced in Machine Learning by Minka (2001). Consider a generic posterior:

$$\pi(\theta) = p(\theta|y) \propto p(\theta) \prod_{i=1}^{n} l_i(\theta)$$  \hfill (2)

where the $l_i$ are $n$ contributions to the likelihood. Aim is to approximate $\pi$ with

$$q(\theta) \propto \prod_{i=0}^{n} f_i(\theta)$$  \hfill (3)

where the $f_i$'s are the “sites”. To obtain a Gaussian approximation, take $f_i(\theta) \propto \exp \left( -\frac{1}{2} \theta^t Q_i \theta + r_i \theta \right)$, so that:

$$q(\theta) \propto \exp \left\{ -\frac{1}{2} \theta^t \left( \sum_{i=0}^{n} Q_i \right) \theta + \left( \sum_{i=0}^{n} r_i \right)^t \theta \right\}$$  \hfill (4)

where $Q_i$ and $r_i$ are the site parameters.
We wish to minimise $KL(\pi\|q)$. To that aim, we update each site $(Q_i, r_i)$ in turn, as follows. Consider the hybrid:

$$h_i(\theta) \propto q_{-i}(\theta)l_i(\theta), \quad q_{-i}(\theta) = \prod_{j \neq i} f_j(\theta)$$

and adjust $(Q_i, r_i)$ so that $KL(h_i\|q)$ is minimal. One may easily prove that this may be done by moment matching, i.e. calculate:

$$\mu_h = \mathbb{E}^{h_i}[\theta], \quad \Sigma_h = \mathbb{E}^{h_i}\left[\theta\theta^T\right] - \mu_i\mu_i^T$$

set $Q_h = \Sigma_h^{-1}, r_h = \Sigma_h^{-1}\mu_h$, then adjust $(Q_i, r_i)$ so that $(Q_h, r_h)$ and $(Q, r) = (\sum_{i=0}^n Q_i, \sum_{i=0}^n r_i)$ (the moments of $q$) match.

$$Q_i \leftarrow \Sigma_h^{-1} - Q_{-i}, \quad r_i \leftarrow \Sigma_h^{-1}\mu_h - r_{-i}.$$
Convergence is usually obtained after a few complete cycles over all the sites.

Output is a Gaussian distribution which is “closest” to target $\pi$, in KL sense.

We use the Gaussian family for $q$, but one may take another exponential family.

Feasibility of EP is determined by how easy it is to compute the moments of order 1 and 2 of the hybrid distribution (i.e. a Gaussian density $q_i$ times a single likelihood contribution $l_i$).
Going back to the EP-ABC target:

\[
p_\epsilon(\theta|y^*) \propto p(\theta) \prod_{i=1}^{n} \left\{ \int p(y_i|y_{1:i-1}^*, \theta) \mathbb{1}_{\{\|y_i-y_i^*\|\leq \varepsilon\}} \, dy_i \right\}
\]

we take

\[
l_i(\theta) = \int p(y_i|y_{1:i-1}^*, \theta) \mathbb{1}_{\{\|y_i-y_i^*\|\leq \varepsilon\}} \, dy_i.
\]

In that case, the hybrid distribution is a Gaussian times \(l_i\). The moments are not available in close-form (obviously), but they are easily obtained, using some form of ABC for a single observation.
Inputs: $\epsilon$, $y^*$, $i$, and the moment parameters $\mu_{-i}$, $\Sigma_{-i}$ of the Gaussian pseudo-prior $q_{-i}$.

1. Draw $M$ variates $\theta^{[m]}$ from a $N(\mu_{-i}, \Sigma_{-i})$ distribution.
2. For each $\theta^{[m]}$, draw $y_i^{[m]} \sim p(y_i|y_{1:i-1}^*, \theta^{[m]})$.
3. Compute the empirical moments

\[
M_{\text{acc}} = \sum_{m=1}^{M} \mathbb{1}_{\{\|y_i^{[m]} - y_i^*\| \leq \epsilon\}}, \quad \hat{\mu}_h = \frac{\sum_{m=1}^{M} \theta^{[m]} \mathbb{1}_{\{\|y_i^{[m]} - y_i^*\| \leq \epsilon\}}}{M_{\text{acc}}}
\]

(6)

\[
\hat{\Sigma}_h = \frac{\sum_{m=1}^{M} \theta^{[m]} \{\theta^{[m]}\}^t \mathbb{1}_{\{\|y_i^{[m]} - y_i^*\| \leq \epsilon\}}}{M_{\text{acc}}} - \hat{\mu}(h_i)\hat{\mu}(h_i)^t. \quad (7)
\]

Return $\hat{Z}(h_i) = M_{\text{acc}}/M$, $\hat{\mu}(h_i)$ and $\hat{\Sigma}(h_i)$. 

Return $\hat{Z}(h_i) = M_{\text{acc}}/M$, $\hat{\mu}(h_i)$ and $\hat{\Sigma}(h_i)$. 

Return $\hat{Z}(h_i) = M_{\text{acc}}/M$, $\hat{\mu}(h_i)$ and $\hat{\Sigma}(h_i)$. 

Return $\hat{Z}(h_i) = M_{\text{acc}}/M$, $\hat{\mu}(h_i)$ and $\hat{\Sigma}(h_i)$.
We are turning a deterministic, fixed-point algorithm, into a stochastic algorithm, hence numerical stability may be an issue. Solutions:

- We adjust dynamically $M$ the number of simulated points at a given site, so that the number of accepted points exceeds some threshold.
- We use Quasi-Monte Carlo in the $\theta$ dimension.
- Slow EP updates may also be used.
In the IID case, \( p(y_i|y_{1:i-1}, \theta) = p(y_i|\theta) \), and the simulation step \( y_i^{[m]} \sim p(y_i|\theta^{[m]}) \) is the same for all the sites, so it is possible to recycle simulations, using importance sampling.
An \textit{IID} univariate model taken from Peters et al. (2010). The observations are alpha-stable, with common distribution defined through the characteristic function

\[ \Phi_X(t) = \begin{cases} 
\exp \left\{ i\delta t - \gamma^\alpha |t|^\alpha \left[ 1 + i\beta \tan \frac{\pi \alpha}{2} \text{sgn}(t)(|\gamma t| - 1) \right] \right\} & \alpha \neq 1 \\
\exp \left\{ i\delta t - \gamma |t| \left[ 1 + i\beta \frac{2}{\pi} \text{sgn}(t) \log |\gamma t| \right] \right\} & \alpha = 1 
\end{cases} \]

Density is not available in close-form.
Data: \( n = 1200 \) AUD/GBP log-returns computed from daily exchange rates.
Marginal posterior distributions of $\alpha$, $\beta$, $\gamma$ and $\delta$ for alpha-stable model: MCMC output from the exact algorithm (histograms, 60h), approximate posteriors provided by EP-ABC (40min, solid line), kernel density estimates computed from MCMC-ABC sample based on summary statistic proposed by Peters et al (50 times more simulations, dashed line).
Second example: Lokta-Volterra processes

The stochastic Lotka-Volterra process describes the evolution of two species $Y_1$ (prey) and $Y_2$ (predator):

$$
\begin{align*}
Y_1 & \xrightarrow{r_1} 2Y_1 \\
Y_1 + Y_2 & \xrightarrow{r_2} 2Y_2 \\
Y_2 & \xrightarrow{r_3} \emptyset
\end{align*}
$$

We take $\theta = (\log r_1, \log r_2, \log r_3)$, and we observe the process at discrete times. Model is Markov, $p(y_i^*|y_{1:i-1}^*, \theta) = p(y_i^*|y_{i-1}^*, \theta)$. 

PMCMC approximations of the ABC target (histograms) for $\epsilon = 3$ (top), EP-ABC approximations, for $\epsilon = 3$ (top) and $\epsilon = 1$ (bottom).
Third example: reaction times

Subject must choose between $k$ alternatives. Evidence $e_j(t)$ in favour of choice $j$ follows a Brownian motion with drift:

$$\tau de_j(t) = m_j dt + dW_t^j.$$  

Decision is taken when one evidence “wins the race”; see plot.
1860 Observations, from a single human being, who must choose between “signal absent”, and “signal present”.

### Relative target contrast vs. Reaction time (ms)

<table>
<thead>
<tr>
<th>Position</th>
<th>Reaction time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>600</td>
</tr>
<tr>
<td></td>
<td>700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>Reaction time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>600</td>
</tr>
<tr>
<td></td>
<td>700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</th>
<th>Reaction time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>400</td>
</tr>
<tr>
<td></td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>600</td>
</tr>
<tr>
<td></td>
<td>700</td>
</tr>
</tbody>
</table>

### Position A

- **Relative target contrast**: 0.05, 0.10, 0.15, 0.20
- **Reaction time (ms)**: 200, 300, 400, 500, 600, 700

### Position B

- **Relative target contrast**: 0.05, 0.10, 0.15, 0.20
- **Reaction time (ms)**: 200, 300, 400, 500, 600, 700

### Position C

- **Relative target contrast**: 0.05, 0.10, 0.15, 0.20
- **Reaction time (ms)**: 200, 300, 400, 500, 600, 700

---

**Note**: The diagrams show scatter plots with linear trends for both “signal absent” and “signal present” responses at different positions and relative target contrasts.
Results

The figure illustrates the effectiveness of the model in predicting reaction times (ms) for different positions (A, B, C) and relative target contrasts (0.05 to 0.20). Each plot shows individual data points (grey circles) and the predictive density (black line) along with the real data (black dots). The 90% quantile (dashed line) and mean (solid line) are also indicated. The plots clearly demonstrate the model's ability to capture the variability in reaction times across different conditions.
Conclusion

- EP-ABC features two levels of approximations: EP, and ABC (\(\varepsilon\), no summary stat.).
- Standard ABC also has two levels of approximations: ABC (\(\varepsilon\)), plus summary stats.
- EP-ABC is fast (minutes), because it integrates one datapoint at a time (not all of them together).
- EP-ABC also approximates the evidence.
- Current scope of EP-ABC is restricted to models such that one may sample from \(p(y_i|y_{1:i-1}^*)\).
- Convergence of EP-ABC is an open problem.
“It seems quite absurd to reject an EP-based approach, if the only alternative is an ABC approach based on summary statistics, which introduces a bias which seems both larger (according to our numerical examples) and more arbitrary, in the sense that in real-world applications one has little intuition and even less mathematical guidance on to why $p(\theta | s(y))$ should be close to $p(\theta | y)$ for a given set of summary statistics.”