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

Nicolas Chopin

CREST (ENSAE)

joint work with Simon Barthelmé (TU Berlin)

Basic ABC

Data: $\boldsymbol{y}^\star$, prior $p(\boldsymbol{\theta})$, model $p(\boldsymbol{y}|\boldsymbol{\theta})$. Likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ is intractable.

1. Sample $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$
2. Sample $\boldsymbol{y} \sim p(\boldsymbol{y}|\boldsymbol{\theta})$
3. Accept $\boldsymbol{\theta}$ iff $\|s(\boldsymbol{y}) - s(\boldsymbol{y}^\star)\| \leq \epsilon$

ABC target

The previous algorithm targets:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \int p(\boldsymbol{y}|\boldsymbol{\theta}) \mathbb{1}_{\{\|s(\boldsymbol{y})-s(\boldsymbol{y}^\star)\|\leq\epsilon\}} \, d\boldsymbol{y}$$

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

1. Non-parametric error, governed by "bandwidth" ϵ; $p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \rightarrow p(\boldsymbol{\theta}|s(\boldsymbol{y}^\star))$ as $\epsilon \rightarrow 0$.
2. Bias introduced by summary stat. s, since $p(\boldsymbol{\theta}|s(\boldsymbol{y}^\star)) \neq p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$.

Note that $p(\boldsymbol{\theta}|s(\boldsymbol{y}^\star)) \approx p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$ may be a reasonable approximation, but $p(\boldsymbol{y}^\star)$ and $p(s(\boldsymbol{y}^\star))$ have no clear relation: hence standard ABC cannot reliably approximate the evidence.

EP-ABC target

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

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y^\star_{1:i-1}, \boldsymbol{\theta}) \mathbb{1}_{\left\{\|y_i - y_i^\star\| \leq \varepsilon\right\}} \, dy_i \right\} \qquad (1)$$

Standard ABC cannot target this approximate posterior, because the probability that $\|y_i - y_i^\star\| \leq \varepsilon$ for all i simultaneously is exponentially small w.r.t. n. But it does not depend on some summary stats s, and $p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \to p(\boldsymbol{\theta}|\boldsymbol{y}^\star)$ 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:

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

where the l_i are n contributions to the likelihood. Aim is to approximate π with

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

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

$$q(\boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{\theta}^t\left(\sum_{i=0}^{n}\mathbf{Q}_i\right)\boldsymbol{\theta} + \left(\sum_{i=0}^{n}\mathbf{r}_i\right)^t\boldsymbol{\theta}\right\} \tag{4}$$

where $\mathbf{Q}_i$ and $\mathbf{r}_i$ are the site parameters.

Site update

We wish to minimise $KL(\pi \| q)$. To that aim, we update each site $(\boldsymbol{Q}_i, \boldsymbol{r}_i)$ in turn, as follows. Consider the hybrid:

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

and adjust $(\boldsymbol{Q}_i, \boldsymbol{r}_i)$ so that $KL(h_i \| q)$ is minimal. One may easily prove that this may be done by moment matching, i.e. calculate:

$$\boldsymbol{\mu}_h = \mathbb{E}^{h_i}\left[\boldsymbol{\theta}\right], \quad \boldsymbol{\Sigma}_h = \mathbb{E}^{h_i}\left[\boldsymbol{\theta}\boldsymbol{\theta}^T\right] - \boldsymbol{\mu}_i \boldsymbol{\mu}_i^T$$

set $\boldsymbol{Q}_h = \boldsymbol{\Sigma}_h^{-1}$, $\boldsymbol{r}_h = \boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h$, then adjust $(\boldsymbol{Q}_i, \boldsymbol{r}_i)$ so that $(\boldsymbol{Q}_h, \boldsymbol{r}_h)$ and $(\boldsymbol{Q}, \boldsymbol{r}) = (\sum_{i=0}^n \boldsymbol{Q}_i, \sum_{i=0}^n \boldsymbol{r}_i)$ (the moments of q) match.

$$\boldsymbol{Q}_i \leftarrow \boldsymbol{\Sigma}_h^{-1} - \boldsymbol{Q}_{-i}, \quad \boldsymbol{r}_i \leftarrow \boldsymbol{\Sigma}_h^{-1}\boldsymbol{\mu}_h - \boldsymbol{r}_{-i}.$$

EP quick summary

- Convergence is usually obtained after a few complete cycles over all the sites.
- Output is a Gaussian distribution which is “closest” to target π, in KL sense.
- We use the Gaussian family for q, but one may take another exponential family.
- Feasiblity 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).

EP-ABC

Going back to the EP-ABC target:

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}^\star) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{n} \left\{ \int p(y_i|y^\star_{1:i-1}, \boldsymbol{\theta}) \mathbb{1}_{\{\|y_i - y_i^\star\| \leq \varepsilon\}} \, dy_i \right\} \qquad (5)$$

we take

$$l_i(\boldsymbol{\theta}) = \int p(y_i|y^\star_{1:i-1}, \boldsymbol{\theta}) \mathbb{1}_{\{\|y_i - y_i^\star\| \leq \varepsilon\}} \, dy_i .$$

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.

EP-ABC site update

Inputs: ϵ, $\boldsymbol{y}^\star$, i, and the moment parameters $\boldsymbol{\mu}_{-i}$, $\boldsymbol{\Sigma}_{-i}$ of the Gaussian pseudo-prior q_{-i}.

1. Draw M variates $\boldsymbol{\theta}^{[m]}$ from a $N(\boldsymbol{\mu}_{-i}, \boldsymbol{\Sigma}_{-i})$ distribution.
2. For each $\boldsymbol{\theta}^{[m]}$, draw $y_i^{[m]} \sim p(y_i | y^\star_{1:i-1}, \boldsymbol{\theta}^{[m]})$.
3. Compute the empirical moments

$$M_{acc} = \sum_{m=1}^{M} \mathbb{1}_{\left\{\|y_i^{[m]} - y_i^\star\| \leq \varepsilon\right\}}, \quad \widehat{\boldsymbol{\mu}}_h = \frac{\sum_{m=1}^{M} \boldsymbol{\theta}^{[m]} \mathbb{1}_{\left\{\|y_i^{[m]} - y_i^\star\| \leq \varepsilon\right\}}}{M_{acc}} \tag{6}$$

$$\widehat{\boldsymbol{\Sigma}}_h = \frac{\sum_{m=1}^{M} \boldsymbol{\theta}^{[m]} \left\{\boldsymbol{\theta}^{[m]}\right\}^t \mathbb{1}_{\left\{\|y_i^{[m]} - y_i^\star\| \leq \varepsilon\right\}}}{M_{acc}} - \widehat{\boldsymbol{\mu}}(h_i)\widehat{\boldsymbol{\mu}}(h_i)^t. \tag{7}$$

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

Numerical stability

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 $\boldsymbol{\theta}$ dimension.
- Slow EP updates may also be used.

Acceleration in the IID case

In the IID case, $p(y_i|y_{1:i-1}, \boldsymbol{\theta}) = p(y_i|\boldsymbol{\theta})$, and the simulation step $y_i^{[m]} \sim p(y_i|\boldsymbol{\theta}^{[m]})$ is the same for all the sites, so it is possible to recycle simulations, using importance sampling.

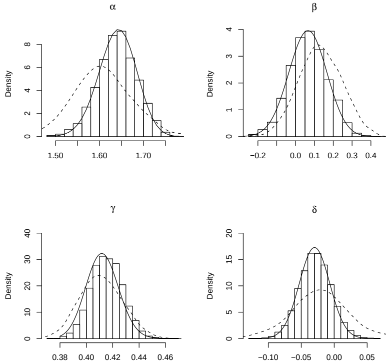
First example: alpha-stable distributions

An IID univariate model taken from Peters et al. (2010). The observations are alpha-stable, with common distribution defined through the characteristic function

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

Density is not available in close-form.
Data: $n = 1200$ AUD/GBP log-returns computed from daily exchange rates.

Results from alpha-stable example



Marginal posterior distributions of α, β, γ and δ 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).

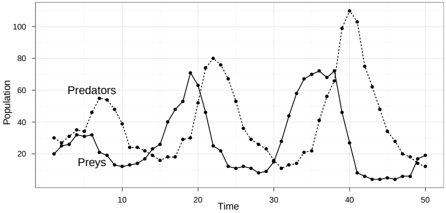
Second example: Lokta-Volterra processes

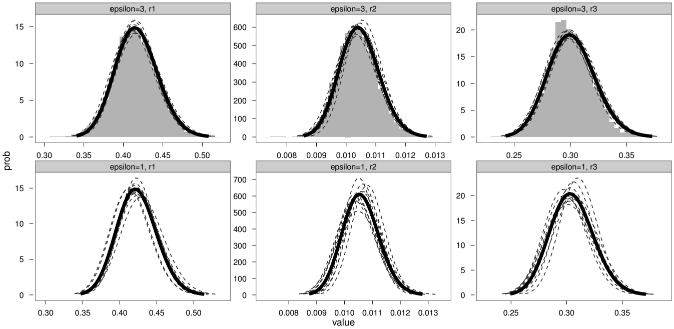
The stochastic Lotka-Volterra process describes the evolution of two species Y_1 (prey) and Y_2 (predator):

$$\begin{aligned} Y_1 &\stackrel{r_1}{\rightarrow} 2Y_1 \\ Y_1 + Y_2 &\stackrel{r_2}{\rightarrow} 2Y_2 \\ Y_2 &\stackrel{r_3}{\rightarrow} \emptyset \end{aligned}$$

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

Simulated data





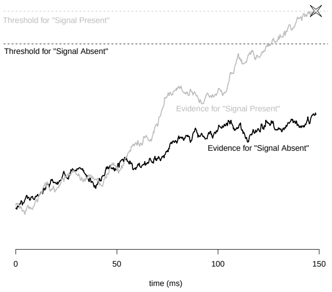
PMCMC approximations of the ABC target (histograms) for $\epsilon = 3$ (top), EP-ABC approximations, for $\epsilon = 3$ (top) and $\epsilon = 1$ (bottom).

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:

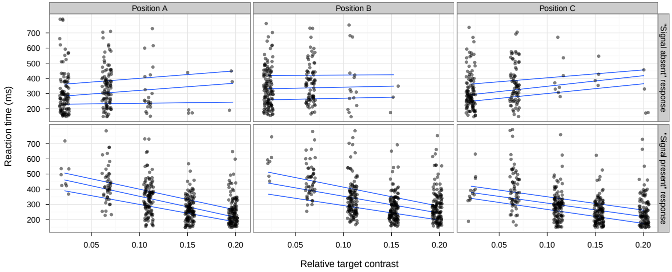
$$\tau de_j(t) = m_j dt + dW_t^j.$$

Decision is taken when one evidence "wins the race"; see plot.

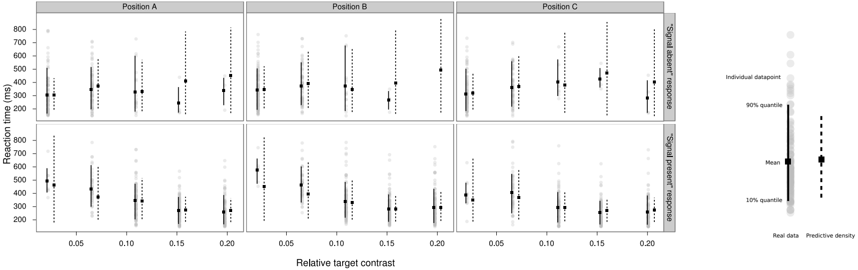


Data

1860 Observations, from a single human being, who must choose between “signal absent”, and “signal present”.



Results


Position A
Position B
Position C
"Signal absent" response
"Signal present" response
Reaction time (ms)
800
700
600
500
400
300
200
0.05
0.10
0.15
0.20
Relative target contrast
Individual datapoint
90% quantile
Mean
10% quantile
Real data
Predictive density

Conclusion

- EP-ABC features two levels of approximations: EP, and ABC (ε, no summary stat.).
- standard ABC also has two levels of approximations: ABC (ε), 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 $p(y_i|y^\star_{1:i-1})$.
- Convergence of EP-ABC is an open problem.

Quote, references

"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 $p(\boldsymbol{\theta}|s(\boldsymbol{y}))$ should be close to $p(\boldsymbol{\theta}|\boldsymbol{y})$ for a given set of summary statistics."

- Barthelmé, S. and Chopin, N. (2011). ABC-EP: Expectation Propagation for Likelihood-free Bayesian Computation, ICML 2011 (Proceedings of the 28th International Conference on Machine Learning), L. Getoor and T. Scheffer (eds), 289-296.
- Barthelmé, S. & Chopin, N. (2011). Expectation-Propagation for Summary-Less, Likelihood-Free Inference, arxiv:1107.5959.