Expectation Propagation for Approximate Bayesian Computation

Felix Held
Fraunhofer-Chalmers Centre and Mathematical Sciences, Chalmers and University of Gothenburg

2018-03-28

Background

- Expectation Propagation (EP) is an algorithm for *variational inference*, i.e. estimation of approximate posterior distribution
- Based on paper Simon Barthelmé & Nicolas Chopin (2014) *Expectation Propagation for Likelihood-Free Inference*, Journal of the American Statistical Association, 109:505, 315-333, DOI: 10.1080/01621459.2013.864178

Setting

- Assume posterior in Bayesian inference can be written as

$$p(\theta|y_{1:N}) \propto p(\theta) \prod_{i=1}^{N} p(y_i|y_{1:i-1}, \theta)$$

- In ABC, we determine

$$p(\theta|y_{1:N}) \propto p(\theta) \prod_{i=1}^{N} \int p(\hat{y}_i|y_{1:i-1}, \theta) \mathbb{1}_{\{\|s_i(\hat{y}_i) - s_i(y_i)\| \leq \varepsilon\}} \mathrm{d}\hat{y}_i$$

  where we assume that we can simulate from $p(\hat{y}_i|y_{1:i-1}, \theta)$. However, no analytic form of the likelihood exists (or expensive to evaluate).

Re-cap: Parametrizations of the normal distribution

- Standard notation with mean μ and covariance matrix Σ

$$\phi(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- Notation in natural parameters: Precision matrix $Q = \Sigma^{-1}$ and precision mean $r = \Sigma^{-1}\mu$

$$\phi(x) \propto \exp\left(-\frac{1}{2}x^T Q x + r^T x\right)$$

Conventional Expectation Propagation

Conventional Expectation Propagation

- Idea: Approximate likelihood factors by some simpler distributions, typically Gaussians. (Originally from (Minka 2013))
- Let here

$$\pi(\theta) \propto \prod_{i=0}^{N} l_i(\theta)$$

where e.g. l_0 is the prior for θ and $l_i = p(y_i | y_{1:i-1}, \theta)$ for $i > 0$.

- We then want to approximate l_i by e.g. a Gaussian

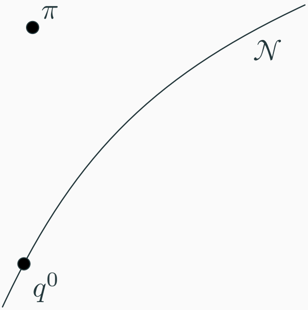
$$f_i(\theta) = \exp\left(-\frac{1}{2}\theta^T Q_i \theta + r_i^T \theta\right)$$

where Q_i is the i-th precision matrix and r_i is the i-th shift.

- Full approximation $q(\theta) \propto \prod_{i=0}^{N} f_i(\theta)$ is then

$$q(x) \propto \exp\left(-\frac{1}{2}\theta^T \left(\sum_{i=0}^{N} Q_i\right)\theta + \left(\sum_{i=0}^{N} r_i\right)^T \theta\right)$$

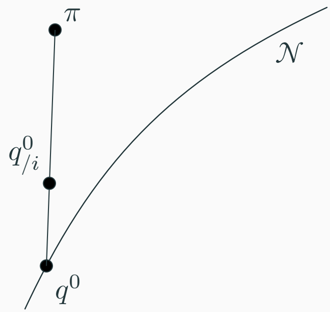
Visual idea



- π: Target distribution
- $\mathcal{N}$: Space of normal distributions
- Move towards target distribution, project back onto space of normal distributions

Visualisation from Simon Barthelmé: The EP algorithm (YouTube)

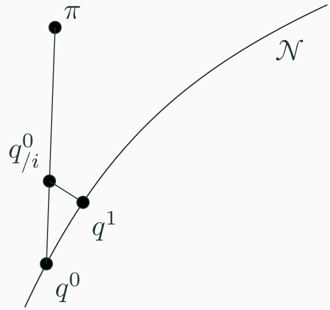
Visual idea



- π: Target distribution
- $\mathcal{N}$: Space of normal distributions
- Move towards target distribution, project back onto space of normal distributions

Visualisation from Simon Barthelmé: The EP algorithm (YouTube)

Visual idea



- π: Target distribution
- $\mathcal{N}$: Space of normal distributions
- Move towards target distribution, project back onto space of normal distributions

Visualisation from Simon Barthelmé: The EP algorithm (YouTube)

Remember

$$\pi(\theta) \propto \prod_{i=0}^{N} l_i(\theta) \quad \text{and} \quad q(\theta) \propto \prod_{i=0}^{N} f_i(\theta)$$

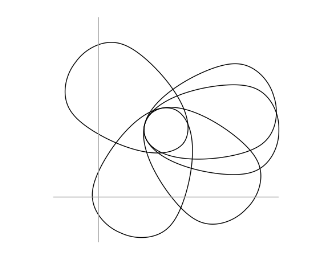
Algorithm

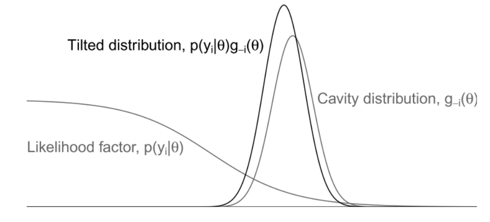
1. Cavity distribution: $q_{-i}(\theta) = \prod_{j \neq i} f_j(\theta)$
2. Hybrid/tilted distribution: $q_{/i}(\theta) = q_{-i}(\theta) l_i(\theta)$
3. Find Gaussian approximation to $q_{/i}(\theta)$ which minimizes Kullback-Leibler divergence

$$\mathrm{KL}(q_{/i} || q^{\mathrm{new}}) = \int q_{/i}(\theta) \log\left(\frac{q_{/i}(\theta)}{q^{\mathrm{new}}(\theta)}\right) \mathrm{d}\theta$$

In the exponential family, this means determination of a new normal approximation by matching the moments of the hybrid distribution.

One factor at a time. Why?





- Cavity acts as a prior for the i-th likelihood factor
- Overlap of likelihood factors is explored more efficiently

Figures from (Gelman et al. 2014)

For Gaussian sites

- We get

$$q_{/i}(\theta) \propto l_i(\theta) q_{-i}(\theta) \propto l_i(\theta) \exp\left(-\frac{1}{2}\theta^T Q_{-i}\theta + r_{-i}^T\theta\right)$$

  where $Q_{-i} = \sum_{j \neq i} Q_j$ and $r_{-i} = \sum_{j \neq i} r_i$.
- Calculate updates

$$Z = \int l_i(\theta) q_{-i}(\theta) \mathrm{d}\theta$$

$$\mu = \frac{1}{Z}\int \theta l_i(\theta) q_{-i}(\theta) \mathrm{d}\theta$$

$$\Sigma = \frac{1}{Z}\int \theta\theta^T l_i(\theta) q_{-i}(\theta) \mathrm{d}\theta - \mu\mu^T$$

- New approximation to $l_i(\theta)$ has parameters

$$Q_i = \Sigma^{-1} - Q_{-i}, \quad r_i = \Sigma^{-1}\mu - r_{-i}$$

Adaptation to likelihood free context

- We do not have access to the analytical form of the likelihood factors $p(\hat{y}_i | y_{1:i-1}, \theta)$
- Integration for moment update not possible analytically
- *Idea:* Sample many

$$\theta^{(m)} \sim q_{-i}(\theta) = \mathcal{N}(\theta; \mu_{-i}, \Sigma_{-i})$$

and sample $\hat{y}_i^{(m)} \sim p(\hat{y}_i | y_{1:i-1}, \theta^{(m)})$ for every $\theta^{(m)}$.

Let $\varepsilon > 0$, $M \in \mathbb{N}$, μ_{-i} and Σ_{-i} be given.

Sample

1. Sample $\theta^{(m)} \sim \mathcal{N}(\theta; \mu_{-i}, \Sigma_{-i})$ for $m = 1, \ldots, M$.
2. Sample $\hat{y}_i^{(m)} \sim p(\hat{y}_i | y_{1:i-1}, \theta^{(m)})$ for every $\theta^{(m)}$.

Compute

$$M_{\text{acc}} = \sum_{m=1}^{M} \mathbb{1}_{\{\|\hat{y}_i^{(m)} - y_i\| \leq \varepsilon\}}$$

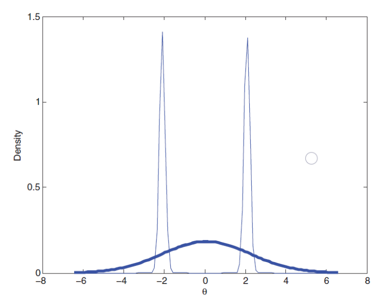
$$\hat{\mu} = \frac{1}{M_{\text{acc}}} \sum_{m=1}^{M} \theta^{(m)} \mathbb{1}_{\{\|\hat{y}_i^{(m)} - y_i\| \leq \varepsilon\}}$$

$$\hat{\Sigma} = \frac{1}{M_{\text{acc}}} \sum_{m=1}^{M} \theta^{(m)} (\theta^{(m)})^T \mathbb{1}_{\{\|\hat{y}_i^{(m)} - y_i\| \leq \varepsilon\}} - \hat{\mu}\hat{\mu}^T$$

Problems and Possible Improvements

Issues

- Multi-modality cannot be captured
- Inversion of covariance matrix is costly and can easily lead to numerical problems (see (Gelman et al. 2014))
  - Σ estimated but $Q = \Sigma^{-1}$ needed for update
- Possibly inefficient ABC scheme, e.g.
  - how many samples?
  - numerical stability?



Possible Improvements

- Enforce minimum number of samples for ABC procedure
- Recycling of already sampled $\theta^{(m)}$ in the case of IID data
  - These can be weighted like

$$w_{i+1}^{(m)} = \frac{q_{-(i+1)}(\theta^{(m)})}{q_{-i}(\theta^{(m)})} \mathbb{1}_{\{\|y^{(m)} - y_{i+1}\| \leq \varepsilon\}}$$

  - Important to monitor the effective sample size

$$\mathrm{ESS} = \frac{\left(\sum_{m=1}^{M} w_i^{(m)}\right)}{\sum_{m=1}^{M} \left(w_i^{(m)}\right)^2}$$

- Quasi-random/low-discrepancy sequences (e.,g. Halton or Sobol' sequences)
- Damping for improved convergence:

$$q_i^{\text{new}}(\theta) = q_i(\theta)^{1-\delta} \left(\hat{q}_{/i}(\theta) \,/\, q_{-i}(\theta)\right)^{\delta}$$

where $\hat{q}_{/i}$ is the Gaussian approximation of $q_{/i}$ and $\delta \in (0, 1]$.

Example: Alpha-stable Models

Alpha-stable Models

- Distributions with characteristic function

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

  where $0 < \alpha \leq 2$, $-1 < \beta < 1$, $\gamma > 0$ and δ are parameters.
- Special cases:
  - $\alpha = 2$ is the normal distribution
  - $\alpha = 1$ is the cauchy distribution
- No closed form density for most α but interesting in e.g. finance and cheap to sample from
- Infinite variance for $\alpha < 2$ and infinite mean for $\alpha < 1$
- It was shown: Hard to determine suitable summary statistics (Peters, Sisson, and Fan 2012)

Numerical Results

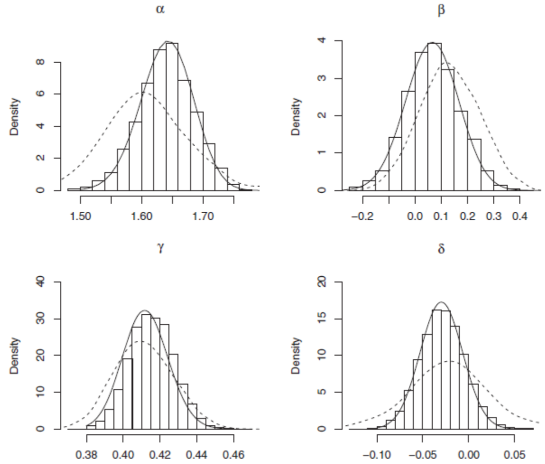


Figure 1. Marginal posterior distributions of α, β, γ, and δ for alpha-stable model: MCMC output from the exact algorithm (histograms), approximate posteriors provided by first run of EP-ABC (solid line), kernel density estimates computed from MCMC-ABC sample based on summary statistic proposed by Peters, Sisson, and Fan (2012) (dashed line).

- $\varepsilon = 0.1$ for EP-ABC and $\varepsilon = 0.03$ for MCMC-ABC

Conclusions

Conclusions

Positive

- Fast and reasonable accurate for unimodal posteriors
- Suitable for distributed computing

Negative

- Gaussian approximation: Might have difficulties for severely skewed posteriors
- Seems to be hard to prove theoretical results

References

Barthelmé, Simon, and Nicolas Chopin. 2014. "Expectation Propagation for Likelihood-Free Inference." *Journal of the American Statistical Association* 109 (505):315–33. https://doi.org/10.1080/01621459.2013.864178.

Gelman, Andrew, Aki Vehtari, Pasi Jylänki, Tuomas Sivula, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, David Schiminovich, and Christian Robert. 2014. "Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data." http://arxiv.org/abs/1412.4869.

Minka, Thomas P. 2013. "Expectation Propagation for approximate Bayesian inference." *Uncertainty in Artificial Intelligence (UAI)* 17 (2):362–69. http://arxiv.org/abs/1301.2294.

Peters, G.W., S.A. Sisson, and Y. Fan. 2012. "Likelihood-Free Bayesian Inference for α-Stable Models." *Computational Statistics & Data Analysis* 56 (11):3743–56. https://doi.org/https://doi.org/10.1016/j.csda.2010.10.004.